(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 December 2001 (13.12.2001)

PCT

(10) International Publication Number
WO 01/95230 A2

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/EP01/06360

(22) International Filing Date: 1 June 2001 (01.06.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/589,167	8 June 2000 (08.06.2000)	US
60/241,844	20 October 2000 (20.10.2000)	US
60/241,801	20 October 2000 (20.10.2000)	US

(71) Applicant (for all designated States except US): VIRCO UK LIMITED [GB/GB]; 184 Cambridge Science Park, Milton Road, Cambridge CB4 0GA (GB).

(72) Inventors; and
(75) Inventors/Applicants (for US only): LARDER, Brendan [GB/GB]; Monona, 6 Churchlane, Cambridge, Cambridgeshire CB3 7HQ (GB). WANG, Dechao [CN/GB]; 21 Claygate Road, Cambridge, Cambridgeshire CB1 9JZ (GB).

(74) Agent: DAELEMANS, Frank; Generaal De Wittelaan L11B3, B-2800 Mechelen (BE).

(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

(54) Title: METHOD AND SYSTEM FOR PREDICTING THERAPEUTIC AGENT RESISTANCE AND FOR DEFINING THE GENETIC BASIS OF DRUG RESISTANCE USING NEURAL NETWORKS

(57) Abstract: A method and system for predicting the resistance of a disease to a therapeutic agent is provided. Further provided is a method and system for designing a therapeutic treatment agent for a patient afflicted with a disease. Specifically, the methods use a trained neural network to interpret genotypic information obtained from the disease. The trained neural network is trained using a database of known or determined genotypic mutations that are correlated with phenotypic therapeutic agent resistance. The present invention also provides methods and systems for predicting the probability of a patient developing a genetic disease. A trained neural network for making such predictions is also provided. Also provided is a method and system for determining the genetic basis of therapeutic agent resistance.

METHOD AND SYSTEM FOR PREDICTING
THERAPEUTIC AGENT RESISTANCE AND FOR DEFINING THE GENETIC
BASIS OF DRUG RESISTANCE USING NEURAL NETWORKS

The present invention relates to methods and systems for predicting the
5   resistance of a disease to a therapeutic agent by application of genotype and phenotype
resistance information in a neural network. The present invention further relates to
methods and systems for designing a therapeutic treatment regimen for a patient based
upon the genotype of the disease afflicting the patient. Under another aspect of the
present invention, methods and systems for predicting the probability that a patient will
10   develop a genetic disease are provided. Under an additional aspect of the present
invention, methods and systems for using neural networks to define the genetic basis of
therapeutic agent resistance are provided. More specifically, the present invention
relates to the use of bioinformatic, molecular biology, and biochemistry tools in such
methods and systems.

15   Since the issuance of the first report suggesting a correlation between the
emergence of viral resistance and clinical progression, techniques to determine the
resistance of a pathogen or malignant cell to a therapeutic agent have been increasingly
incorporated into clinical studies of therapeutic regiments. Brendan Larder *et al.*, HIV
Resistance and Implications for Therapy (1998), herein incorporated by reference.
20   However, the complexity of therapeutic agent resistance makes it difficult to determine
or accurately predict therapeutic agent resistance. With more drugs and therapeutic
options becoming available, therapeutic agent resistance testing is expected to play an
important role in the management and treatment of pathogen infection or cancer.

All of these methods employ two general approaches for measuring resistance
25   to therapeutic agents, namely phenotypic testing and genotypic testing. Phenotypic
testing directly measures the actual sensitivity of a patient's pathogen or malignant cell
to particular therapeutic agents, while genotypic resistance testing examines the
presence of specific genetic mutations or patterns of mutations in the pathogen or
malignant cell that confer resistance to a certain therapeutic agent(s). Although
30   phenotypic testing is believed to be a more comprehensive and accurate assessment of
therapeutic agent resistance than genotypic testing, phenotypic testing can take longer
and is generally more expensive than genotypic testing. Compared with phenotypic
testing, genotypic testing has advantages, including the relative simplicity, low cost,
and the speed with which the test can be performed. However, at present, it remains
35   difficult to interpret the results of a genotypic test to provide meaningful conclusions
about therapeutic agent resistance. See, e.g., Tim Horn and Spencer Cox, *A No-*

Nonsense Guide to HIV Drug Resistance Testing, (Ed. Douglas Richman, M.D., University of California, San Diego.

A number of different approaches are presently available to aid in the interpretation of genotypic testing, including:

A. Interpretation by the Physician

5

A physician can interpret and make a judgement as to the optimum treatment based on knowledge of the primary resistance mutations associated with each therapeutic agent and the patient's recent treatment history. To assist physicians to make these judgements, various expert opinion-panels have been convened and have

10   published guidelines. For example, the Resistance Collaborative Group has published such guidelines for HIV-1. See, e.g., Carpenter *et al., JAMA* 283(3):381-390 (2000), herein incorporated by reference. Obviously, this type of method is highly subjective.

B. Rules-based Algorithms

Rules-based algorithms are essentially a formalized version of the above-

15   identified interpretation method with tables giving the mutations that are associated with resistance to each of the therapeutic agents. These can be simple printed tables or the information can be used to develop a rules-based computer algorithm. An example of such an interpretation system is the VircoGEN™ I system (available from Virco) and the techniques disclosed in WO 97/27480.

C. Statistical Analysis

20

Statistical analyses have been used to compare and relate phenotypes and genotypes. Harrigan *et al.,* "Drug resistance and short term virological response in patients prescribed multidrug rescue therapy,"; Hammer *et al.,* "Relationship of phenotypic and genotypic resistance profiles to virological outcome in a trial of

25   abacavir, nelfinavir, efavirenz and adefovir dipivoxil in patients with virological failure receiving indinavir (ACTG 372),"; Zolopa *et al.,* "A comparison of phenotypic, genotypic and clinical/treatment history predictors of virological response to saquinavir/ritonavir salvage therapy in a clinic-based cohort,"; Vingerhotes *et al.,* "The accuracy and reproducibility of high throughput genotypic and phenotypic HIV-1

30   resistance testing under EN45001 and CLIA accreditation labels,"; Anton *et al.,* "Comparative patterns of HIV-1 genotypic and phenotypic resistance profiles in gut and plasma,"; Hertogs *et al.,* "A blinded comparative analysis of two genotyping service laboratories: full sequence analysis of HIV-1 protease and reverse transcriptase," all presented at the 3rd International Workshop on HIV Drug Resistance

35   & Treatment Strategies, San Diego, USA, 23-26 June 1999, all of which are herein incorporated by reference. These methods provide information about whether phenotypic data correlate to the corresponding genotypes. The faced difficulties are,

however, in relating quantitatively the genotype of any specific sample to its phenotype. Interpreting HIV-1 drug resistance mutation patterns has been improved by predicting the phenotype using a large phenotype-genotype database. To relate a "test" genotype to phenotypic resistance information, a series of genotypic patterns were related to specific drugs. These patterns are attached to all genotypic samples in the database thus enabling rapid searches to be performed. The phenotypes of samples in the database that match a particular genotype can then be retrieved and displayed as the proportion resistant or sensitive to each drug. Larder *et al.*, Predicting HIV-1 phenotypic resistance from genotype using a large phenotype-genotype relational database, 3rd International Workshop on HIV Drug Resistance & Treatment Strategies, San Diego, USA, 23-26 June 1999, herein incorporated by reference. This system makes it possible to obtain a rapid indication of the likely phenotype of a genotyped sample by matching substantial archived phenotypic data to a mutation pattern.

However, little is known about the functional form of the relationship between genotype and phenotype, therefore, making it difficult to utilize parametric modeling approaches. Furthermore, non-independent mutations in genotypic mutation patterns may be involved. This makes it difficult to apply conventional methods to perform function mapping between mutation patterns and the degree of drug resistance.

Currently, however, there are improved relational databases that utilize pattern recognition and phenotypic matching, which have demonstrated a greater than 90% accuracy in predicting phenotypic resistance. Pattern recognition and phenotype matching systems are implemented through software and use the mutations found in the patient sample to search for matches in a database of genotypes and phenotypes from thousands of samples. A search engine is used to scan a phenotype-genotype database. The phenotypes of samples in the database that match a particular genotype can then be retrieved and displayed as the proportion resistant or sensitive to each therapeutic agent. This type of system makes it possible to obtain a rapid indication of the likely phenotype of a genotyped sample by matching substantial archived phenotypic data to a mutation pattern. An example of such a software system is the VirtualPhenotype™ (PCT/EP01/04445)

The present invention provides the next generation of software implemented pattern recognition and phenotype matching systems and employs a neural network to accurately predict the development of therapeutic agent resistance or sensitivity based upon genotypic and phenotypic information and to accurately define the genetic basis of therapeutic agent resistance. Neural networks have been successfully used as pattern classifiers in many applications. See, e.g., Christopher M. Bishop, "Neural Networks for Pattern Recognition," Clarendon Press, Oxford, (1995); Sbirrazzuoli and Brunel, *Neural Comput & Applic.* 5:20-32 (1997); Chow and Cho, *Neural Comput & Applic.*

5:66-75 (1997), the disclosures of which are expressly incorporated herein by reference in their entireties. Until now, however, neural networks have not been used to predict therapeutic agent resistance or sensitivity. With respect to relational databases and approaches like virtual phenotyping, neural networks may provide advantages as to the

5   number of samples required for an accurate analysis, advantages as to the calculation time and advantages in predicting the resistance profiles of drugs not having particular signature mutations.

        To achieve these and other advantages, and in accordance with the principles of the present invention as embodied and broadly described herein, the present invention,

10   in one aspect, provides a method and system for predicting therapeutic agent resistance using a neural network. According to one aspect, the present invention provides a method for predicting resistance of a pathogen to a therapeutic agent comprising: (a) providing a trained neural network; (b) providing a determined genetic sequence from the pathogen; and (c) predicting resistance of the pathogen to the therapeutic agent

15   using the determined genetic sequence and the trained neural network.

        The present invention further provides a method for predicting resistance of a disease to a therapeutic agent comprising: (a) providing a trained neural network; (b) providing a determined genetic sequence from the disease; and (c) predicting resistance of the disease to the therapeutic agent using the determined genetic sequence and the

20   trained neural network.

        Further provided in the present invention is a method for predicting resistance of a pathogen to a therapeutic agent comprising: (a) providing a neural network; (b) training the neural network on a training data set, wherein each member of the training data set corresponds to a genetic mutation that correlates to a change in therapeutic

25   agent resistance; (c) providing a determined genetic sequence from the pathogen; and (d) predicting resistance of the pathogen to the therapeutic agent using the determined genetic sequence of the pathogen and the trained neural network.

        The present invention also provides a trained neural network capable of predicting resistance of a disease to a therapeutic agent, wherein the trained neural

30   network comprises: (a) a set of input nodes, wherein each member of the set of input nodes corresponds to a mutation in the genome of the disease; and (b) a set of output nodes, wherein each member of the set of output nodes corresponds to a therapeutic agent used to treat the disease.

        In another embodiment, the present invention provides a method of designing a

35   therapeutic agent treatment regimen for a patient afflicted with a disease comprising: (a) providing a determined genetic sequence from the disease; (b) inputting the determined genetic sequence into a trained neural network; (c) predicting resistance of the disease to a therapeutic agent using the determined genetic sequence and the trained

neural network; and (d) using the predicted drug resistance to design a therapeutic drug treatment regimen to treat the patient afflicted with the disease.

Under a further embodiment, the present invention provides a method of predicting the probability of a patient developing a genetic disease comprising: (a) providing a trained neural network; (b) providing a determined genetic sequence from a patient sample; and (c) determining the probability of the patient of developing the genetic disease using the determined genetic sequence and the trained neural network.

Another embodiment of the present invention provides a method for identifying a new mutation that confers resistance to a therapeutic agent comprising: (a) providing a first trained neural network, wherein the number of input nodes for the first trained neural network is equal to the number of mutations known to confer therapeutic resistance to a therapeutic agent; (b) providing a second trained neural network, wherein the number of input nodes of the second trained neural network comprises the number of mutations known to confer therapeutic resistance to a therapeutic agent plus at least one; (c) providing a test data set; (d) inputting the test data set into the first and second trained neural networks; (e) comparing the output of the first and second trained neural networks to determine whether the additional mutation confers therapeutic drug resistance to a disease. The above embodiment serves as an example, it should be appreciated that the network architecture does not necessarily requires two different neural networks for the identification of the novel mutations or mutational profiles.

Another embodiment of the present invention provides a method for studying therapeutic agent resistance comprising: (a) mutating a wild-type gene to create a mutant containing a mutation identified using the neural networks of the present invention; (b) culturing the mutant gene in the presence of a therapeutic agent; (c) culturing wild-type gene in the presence of the therapeutic agent; (d) comparing the growth of the mutant gene against the growth of the wild-type gene.

Another embodiment of the present invention provides a method for studying therapeutic agent resistance comprising: (a) mutating a wild-type virus or bacterium to create a mutant virus or bacterium containing a mutation identified using the neural networks of the present invention; (b) culturing the mutant virus or bacterium in the presence of a therapeutic agent; (c) culturing wild-type virus or bacterium in the presence of the therapeutic agent; (d) comparing the growth of the mutant virus or bacterium against the growth of the wild-type virus or bacterium.

In a further embodiment, the invention provides a computer-readable medium containing instructions for causing a computer to perform a method for predicting resistance of a disease to a therapeutic agent using a trained neural network, the method comprising: receiving at least one determined genetic sequence from the disease; and

predicting resistance of the disease to the therapeutic agent using the at least one determined genetic sequence and the trained neural network.

The invention also provides a computer-readable medium containing a set of program instructions for causing a computer to provide a neural network to perform a
5   method for predicting resistance of a disease to a therapeutic agent, the set of program instructions comprising: means for receiving at least one determined genetic sequence from the disease; and means for predicting resistance of the disease to the therapeutic agent using the at least one determined genetic sequence and the trained neural network.
10

It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention, as claimed. Further features and/or variations may be provided in addition to those set forth herein. For example, the present invention may
15   be directed to various combinations and subcombinations of the disclosed features and/or combinations and subcombinations of several further features disclosed below in the detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of
20   this specification, illustrate various embodiments and/or features of the invention and · together with the description, serve to explain the principles of the invention. In the drawings:

Figure 1 depicts an exemplary framework for capturing the relationship between genotype and phenotypic resistance;
25   Figure 2 depicts an exemplary flowchart for predicting phenotypic resistance based upon genotypic information using a neural network in accordance with the present invention;

Figure 3 depicts an exemplary framework for a three-layer neural network. This exemplary network has I inputs, J hidden units and K output units, and two bias
30   units, both of which have an input signal of 1 (i.e., x_0 and z_0). This exemplary three-layer neural network also has two layers of adaptive weights (w_{ji} and w_{jk}), which are the weight of the jth hidden unit associated with input signal x_i, and the weight of the kth output unit associated with the hidden signal z_j, respectively;

Figure 4(a) is an exemplary comparison between the training and testing errors
35   against the number of hidden nodes;

Figure 4(b) is an exemplary comparison between the number of training and testing errors against the error tolerance index;

Figure 5 is an exemplary plot of the magnitude of resistance for twelve mutation patterns;

Figure 6 is an illustrative graph of the concordance rate between PI genotypes and phenotypes from a neural network with noisy data involved in the training set; and

Figure 7 is an illustrative graph of the concordance rate from a neural network without noisy data involved in the training set.

Figure 8 provides a regression analysis between the predicted phenotypes and the actual phenotypes using 30 mutations.

Figure 9 provides a regression analysis between predicted phenotypes and the actual phenotypes using 90 mutations/polymorphisms.

Figure 10 provides an exemplary framework for identifying new genotype (genos) mutations that confer phenotypic (phenos) therapeutic agent resistance. In this framework, x and y refer to the number of mutations being studied by the neural network, x-VP refers to the "virtual" phenotype for the x-mutation model, and x-DPVP refers to the difference between the "real" phenotype and the "virtual" phenotype.

Figure 11 is an illustrative bar chart of the mutations that improved the prediction of the neural network from the 9- to 26-model.

Figure 12 is an illustrative bar chart of the mutations that improved the prediction of the neural network from the 9- to 60-model.

DETAILED DESCRIPTION OF THE INVENTION

Over time, many patients experience treatment failure or reduced efficacy. In many instances, this is due to mutations in the genome of the pathogen or diseased cell such as a malignant or inflammatory cell, which results in the development of resistance to a therapeutic agent. In other instances, selection by the therapeutic agent results in the accumulation or propagation of variants of the pathogen or malignant cell that had pre-existing resistance to the therapeutic agent. Accordingly, there is a need to monitor a patient's disease state and alter the therapeutic regimen when treatment failure or reduced efficacy occurs. As used herein, the term "disease" and "disease-causing agent" both refer to a nucleic acid, a protein, a pathogen or diseased cell such as, for example, a malignant cell, proliferative cell, inflammatory cell, or any mutated cell, such as a mutated neural cell, that causes, for example, a pathological condition in an organism from the pathogen's infection or malignant cell's replication.

The present invention describes a generic framework for predicting the resistance of a pathogen or malignant cell to a therapeutic agent. The generic frame-work of the present invention can be further used to identify mutation(s) or mutation patterns, including insertions and deletions, that confer resistance to a therapeutic agent. It is understood that the use of the term mutation also includes genetic

-8-

polymorphisms. By employing bioinformatic tools to genotyping and phenotyping methodologies, the present invention accurately predicts resistance of patient's pathogen or malignant cell to a therapeutic agent based on genotypic mutations in the pathogen or malignant cell. First, the disease for which therapeutic resistance is to be

5   predicted is selected. After the selection of the disease, a genotype-phenotype database of therapeutic resistance is located or created. Using this information, the neural network is configured and trained. With the trained network, it is possible to predict therapeutic agent resistance based upon genetic information from the patient's disease.

In one embodiment of the present invention expression levels of proteins or

10   nucleic acids are used

In view of the breadth of the application and the possible use of it for resistance testing, drug profiling, diagnosis, different forms of mutations should be described. Those types of mutations should encompass both genetic and epigenetic mutations. The genetic changes encompasses, (i) base substitutions such as single nucleotide

15   polymorphisms, transitions, transversions, substitutions and (ii) frame shift mutations such as insertions, repeats and deletions. Further to this also microsatellites are useful for the practice of the instant invention. The influence of mutations on the etiology of cancer can be exemplified by the mutations influencing the effect of the tumor suppressor gene p53 (other tumor suppressor genes are TGF-beta, NF-1, WT-1, Rb).

20   Alternatively, mutations present in oncogenes (an example of an oncogene is Ras, other oncogenes are c-myc, c-raf, neu, IL-2), repair genes (e.g. methylguanosyl methyltransferase can cause changes in the phenotype and/or drug effect.

Epigenetic changes encompass alterations of nucleic acids e.g. methylation of nucleic acids. The role of methylation in disease and health has recently been shown by

25   the influence of methylation in different syndromes such as Fragile X and Rett syndrome. It should be understood that methylation not only has an impact on disease status but also on drug profiling (Esteller M. et al. New England Journal of Medicine, 2000, Vol 343:19, p. 1350-1354. "Inactivation of the DNA repair gene MGMT and the clinical response of gliomas to alkylating drugs").

30   Under another embodiment, the generic framework of the present invention can be used to predict the development of a genetic disease in a patient. As used herein, the term "genetic disease" refers to any pathological condition that is directly or indirectly correlated to a genetic mutation. Under this embodiment, a phenotype-genotype database of genetic mutations correlated to with the development of a genetic disease is

35   either located or generated. Using the data from this database, a neural network is trained. A sample from the patient's genetic information is genotyped. By inputting the patient's determined genetic information into the trained neural network, a prediction may be made as to the probability of the patient developing a given disease.

-9-

Using this embodiment of the present invention, the probability of developing any genetic disease associated with a genetic mutation can be determined.

Accordingly, the present invention represents a paradigm shift in the ability of the clinician to monitor a patient's disease state and to accurately prescribe a
5   therapeutic agent or combination of therapeutic agents based upon the pathogen's or malignant cell's existing or developed therapeutic agent resistance, and thereby most effectively treat the patient's disease state.

The present invention can predict the therapeutic agent resistance of any pathogen or malignant cell provided the target sequence is known. A pathogen, as used
10   herein, refers to any disease-producing microorganism, including bacteria, viruses, algae, fungi, yeast and protozoa. A malignant cell, as used herein, refers to a cell having the properties of anaplasia, invasion and metastasis.

The present invention has particular application to the prediction of therapeutic agent resistance of a disease-producing virus. Specifically, the present invention can
15   predict the resistance of human immunodeficiency virus (HIV) type 1 and 2, herpes simplex virus (HSV) type 1 and 2, human papillomavirus virus, hepatitis B virus (HBV), hepatitis C virus (HCV), cytomegalovirus (CMV), rous sarcoma virus (RSV) and Epstein-Barr virus (EBV). The present invention has further particular application to the prediction of therapeutic agent resistance in such disease-producing bacteria as
20   mycobacterium sp., salmonella sp., eschericia sp. and streptococcus sp.

Although some treatment regimens employ a single therapeutic agent, it is more typical today to employ a combination of therapeutic agents to treat any given disease-state. A therapeutic agent, as used herein, refers to any animal, vegetable, mineral, or pharmaceutical substance used to treat a pathogen or malignant cell. It is understood
25   that the term "pharmaceutical substance" refers to pharmaceutical compounds, proteins, DNAs (including anti-sense DNA), RNAs. It should be understood that the nucleic and peptidic compounds can contain non-natural amino acids or bases, known to the one skilled in the art. In addition the nucleic and peptidic compounds can contain non-natural base linkages or peptides bonds known in the art. When a combination of
30   therapeutic agents are employed and resistance develops, the clinician often does not know which therapeutic agent is no longer effective to treat the disease-state. Therapeutic agent resistance can be pre-existing or developed by prolonged exposure to one or more therapeutic agents. Therefore it should be understood that a therapeutic agent comprises also combinations of different compounds.
35   The development of therapeutic agent resistance is especially troubling because, even today, a clinician only has a limited number of therapeutic agents available to treat any given pathogen or malignant cell. Thus, the clinician cannot simply replace all of the therapeutic agents presently being administered with a new set of therapeutic

agents. For example, by replacing the current treatment regimen with a completely new treatment regimen, the clinician may discard an effective therapeutic agent. The clinician also cannot sequentially replace each therapeutic agent being administered in a combination therapy. Moreover, it is not uncommon for a pathogen or malignant cell,

5   which displays resistance to a particular therapeutic agent to also display varying degrees of cross-resistance to other therapeutic agents.

Furthermore, not every mutation causes resistance. For example, a mutation can cause a pathogen or malignant cell to become more sensitive to a given therapeutic agent. Also a mutation can restore drug sensitivity to a pathogen or malignant cell that

10   was previously resistant to that therapeutic agent.

By continual monitoring of the disease-state, the clinician will also be able to assess whether a more effective therapeutic agent can be prescribed to treat the patient. It is understood that the present invention can be equally used to detect the development of therapeutic agent sensitivity in a pathogen or malignant cell. By the

15   use of the term therapeutic agent resistance, it is understood that this term includes both the increase and decrease in the sensitivity of the pathogen or malignant cell to a therapeutic agent.

Therefore, the present invention has particular application to monitoring the effectiveness of combination therapeutic agent treatment regimens. By monitoring the

20   genotypic information of the pathogen or malignant cell, the clinician is able to more accurately assess the effectiveness of the present treatment regimen and prescribe the appropriate replacement therapeutic agent(s) as resistance or sensitivity develops.

Although the present invention is often stated in terms of the treatment of a human patient, it is understood that the present invention can be applied to measure the

25   therapeutic agent resistance of a pathogen or malignant cell that causes a disease state in any animal.

Once new mutation(s) or mutation patterns have been identified using the neural networks of the present invention, a skilled practitioner can construct mutant forms of the wild-type disease. The skilled practitioner can then use the mutant form of the

30   wild-type disease to study therapeutic agent resistance. Although in no way limiting the present embodiment, as an example, the skilled practitioner can perform site directed mutagenesis on a wild-type strain of a virus or bacterium that is normally sensitive to a therapeutic agent to study the effect of that mutation on therapeutic agent resistance. The site directed mutagenesis would be used to create at least one of the

35   mutations identified using the neural networks of the present invention in a wild-type virus or bacterium. The mutant virus or bacterium would then be cultured in the presence of a therapeutic agent and the growth of that mutant virus or bacterium would be measured against the growth of the wild-type virus or bacterium in the presence of

-11-

the therapeutic agent. Any difference in growth rates could then be attributed to the mutation identified by the neural network.

A. Genotyping Methodologies

Genotyping methodologies detect specific genetic changes or mutations, including insertions and deletions, in the genetic information of the sample. Under one embodiment of the present invention, the genotyping methodologies are used to detect specific genetic changes or mutations, including insertions and deletions, in a pathogen or malignant cell genome that are known to be associated with therapeutic agent resistance. As used herein, the term "genome" is meant to refer to any DNA or RNA isolated from the pathogen or malignant cell. Thus, the term genome includes, for example, chromosomal DNA, extra-chromosomal DNA (including plasmid DNA, microsatellite DNA, and mitochondrial DNA), messenger RNA (mRNA), virally encoded DNA or RNA, and the like. These mutations can either make the pathogen or malignant cell more sensitive or more resistant to a therapeutic agent.

Under another embodiment of the present invention, the genotyping methodologies are used to detect specific genetic changes or mutations, including insertions and deletions, in a patient's genome. Preferably, the genotyping methodologies of the present invention are used to detect mutations correlated with the development of a genetic disease. It is understood that not every mutation is directly correlated with a genetic disease. Sickle cell anemia is an example of a genetic mutation that is directly correlated with a genetic disease. Most mutations, however, are indirectly correlated with a genetic disease. These mutations generally increase the prevalence of a patient developing the genetic disease associated with the mutation, but the presence of the mutation, in and of itself, is not determinative of the development of the genetic disease. It is understood that the present invention has particular application to the prediction of the development of a genetic disease that is indirectly correlated to a mutation(s).

Genotyping is simpler to conduct than phenotyping and less expensive. One disadvantage is that the results are difficult to interpret. It is important to note that genotyping is not a measure of resistance on its own – resistance can be inferred from genotypic information but this requires sophisticated interpretation using such methods as described in the present invention.

The interpretation of genotypic is difficult and requires a sophisticated approach. Statistical methods suffer from decreasing accuracy if the problem to be solved, i.c. the relation between mutation(s) and drug efficacy, becomes complex. Such a problem is often encountered where complex genotype patterns are linked to monotherapies or combination therapies as is the case during treatment of infectious

-12-

diseases and malignancies for example. The complex relations between genotypic profiles found in the disease-causing agent, either upon treatment or even under treatment naïve conditions, and the possible therapies can be approached by neural networks. A neural network enables the calculation of resistance more accurately than conventional statistical analyses.

It is understood that any method capable of detecting genetic changes can be used in the present invention. Moreover, these genetic changes can be detected in any DNA or RNA isolated from the sample. In addition, the genetic changes can be detected in cDNA prepared from the sample.

It is understood that the sample may be obtained from an individual of the species that is to be analyzed using any of a number of "invasive" or "non-invasive" sampling means. A sampling means is said to be "invasive" if it involves the collection of the sample from within the skin or organs of a patient such as blood collection, semen collection, needle biopsy, pleural aspiration, etc. In contrast, a "non-invasive" sampling means is one in which the sample is recovered from an internal or external surface of the patient such as swabbing, collection of tears, saliva, urine, fecal material, sweat or perspiration, ductal lavage, etc.

Under one embodiment of the present invention, the DNA or RNA from the pathogen or malignant cell contained in the sample is isolated after the sample has been collected. Techniques for isolating DNA or RNA from a patient sample are known to persons of skill in the art and are fully described in Sambrook et al., Molecular Cloning: A Laboratory Manual, Vols. 1-3, 2nd ed., Cold Spring Harbor Laboratory Press (1989), herein incorporated by reference. The genotypic information is then determined from the isolated DNA or RNA. Alternatively, the genotypic information can be determined directly from the pathogen or malignant cell contained in the sample. A determined genetic sequence as used herein refers to any DNA or RNA from the sample whose sequence has been determined, in whole or in part, preferably using one of the genotyping methodologies of the present invention. Two preferred methodologies suitable for determining genetic sequence are hybridization-based point mutation assays and DNA sequencing.

Hybridization-based point mutation assays search for individual known mutations. While these methods are highly specific, the point mutation assays are reported to only detect a fraction of these known mutations. See, e.g., Stuyver *et al.*, *Antimicrob. Agents Chemotherap. 41*:284-291 (1997) andcan, therefore, only provide a small select part of the resistance picture. Common point mutation assays suitable for use in the present invention include, but are not limited to, primer-specific polymerase chain reaction (PCR) (see, e.g., Larder *et al.*, *AIDS 5*: 137-144 (1991); differential hybridization (see, e.g., Eastman *et al.*, *J. Acquir. Immune Defic. Syndr. Human*

-13-

Retrovirol. 9: 264-273 (1995); Line Probe Assay (LiPA™, Innogenetics) (Larder *et al.*, *AIDS 5*: 137-144 (1991); Stuyver *et al.*, *Antimicrob. Agents Chemother. 41(2)*:284-91 (1997), and gene chip sequencing (see, e.g., Fodor, *Nature 227*:393-395 (1997); U.S. Patent Nos. 5,925,525, 5,861,242, and 5, 834,758). All these references are expressly incorporated herein by reference. Other assays to determine mutations have become available *e.g.* Invader® assay (Third Wave Technologies, Inc.), WAVE® DNA assay (Transgenomic, Inc.), mass spectrometry (Jackson P., et al. *Molecular Medicine Today* 6, 271-276, (2000)) and surface plasmon resonance (Nakatani, K. et al. *Nature Biotechnology* 19(1), 18-19, (2001). An overview of currently used mutation techniques, comprising gel based and non-gel based analyses are surveyed in Shi, M. *Clin. Chem.* 2001, (47:2) 164-172.

DNA sequencing provides information on all the nucleotides in the region of the RNA or DNA sequenced. There are two main types of DNA sequencing methods, the so-called chain-termination method and chemical sequencing (see, e.g., Sanger and Coulson, *J. Mol. Bio. 94*:441-448 (1975), Maxam and Gilbert, *Methods Enzymol. 65*:499-560 (1980), both expressly incorporated herein by reference). Alternative sequencing strategies have been developed e.g. mass spectrometric analysis. Preferably, only a segment or portion of the genetic information from the sample is used to detect a mutation. However, it is understood that the entire genome of a sample can be used to detect a mutation. As used herein, the phrase "mutation" refers to a specific genetic change in the nucleotide sequence of the sample in comparison to the genetic sequence at the same position or location in the wild-type sample, including but not limited to insertions and deletions. The genetic mutation is normally written as in reference to the wild type, *i.e.*, K101N refers to replacement of a Lysine at codon 101 with an Asparagine. However, the mutations of the invention do not depend on the wild-type example listed in order to be within the practice of the invention. For example, the mutation 101N, refers to an Asparagine at the 101 codon regardless of the whether there was a Lysine at 101 prior to mutation.

Under one embodiment of the present invention, it is preferred to select a segment or portion of the genetic information that is known or believed to accumulate mutations that effect drug resistance. Under another embodiment, it is preferred to select a segment or portion of the genetic information from the patient sample that is known or believed to accumulate mutations correlated with the development of a disease. Generally, these segments are genes or fragments of genes encoding enzymes or proteins. Generally, these proteins are associated with the cell membrane. For example, in HIV, genes known to accumulate mutations that effect drug resistance include for example the polymerase gene, the protease gene, envelope protein and the

reverse transcriptase gene. Additional HIV genes of interest for the practice of the current invention are e.g. TAT, ref, nef, integrase gp41, gp120, gp160. From hepatitis B the following genes such as DNA polymerase core. From hepatitis C genes as core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B. From tumor biology different

5   genes linked to disease state have already been identified such as HER2, EGF-receptor, raf, p53, Bcr-Abl, Bcl2 and APC. The effect of mutations in the genes is described for e.g. raf, p53, Bcl2 and APC

Dependent on which method is used, some or all of the mutations that have occurred will be identified. However, the prediction of what the net effect of these

10   mutations might be on the susceptibility of the pathogen or malignant cell population to the various therapeutic agents requires sophisticated interpretation. For example, extensive genetic analysis of resistant viral isolates generated through *in vivo* or *in vitro* selection has revealed that resistance is generally caused by mutations altering the nucleotide sequence at some specific site(s) of the genome. It is then up to the

15   physician to combine this information with all the other information relating to the patient and decide what all this means in terms of selecting drugs for the treatment of their individual patient.

The mutational patterns that have been observed and reported for HIV-1 and that are correlated with drug resistance are very diverse: some antiretroviral agents require only

20   one single genetic change, while others require multiple mutations for resistance to appear. In HIV-1 there are currently approximately 100 mutations that are thought to be involved in the development of HIV-1 therapeutic agent resistance. One such example is N88S, which causes *in vitro* hypersensitivity to amprenavir. See, e.g., Ziermann *et al.*, *J. Virol.* 74(9):4414-9 (2000). A summary of mutations in the HIV

25   genome correlated with drug resistance has been reported. Schinazi *et al.*, *Int. Antiviral News.* 5:129-142 (1997), herein incorporated by reference. Additionally, an electronic listing with mutations has also become available at http://hiv-web.lanl.gov; http://hivdb.stanford.edu/hiv/; or http://www.viral-resistance.com.

30   The relationship between these point mutations, deletions and insertions and the actual susceptibility of HIV-1 to therapeutic agents is extremely complex and interactive. For example, the M184V mutation in HIV-1 confers resistance to 3TC reverses AZT resistance. See, e.g., Larder *et al.*, *Science* 269:696-699 (1995), expressly incorporated herein by reference. The 333D/E mutation, however, reverses

35   this effect and can lead to dual AZT/3TC resistance. See, e.g., Kemp *et al.*, *J. Virol.* 72(6):5093-5098 (1998), expressly incorporated herein by reference.

When HIV-1 is genotyped, the preferred method for genotyping is the VircoGEN™ genotypic test (Virco). The VircoGEN™ is a genotyping assay that uses

sequencing technology to identify all the known resistance mutations that have occurred in the protease - reverse transcriptase (PR-RT) genes of a patient's HIV-1 virus population. This is an indirect measure based on genetic indicators of therapeutic agent resistance.

5   The interpretation of genotypic data is both complex and critically important. As more therapeutic agents are developed and more mutations are correlated to the development of therapeutic agent resistance, this complexity will increase.

B. Phenotyping Methodologies

Phenotyping methodologies measure the ability of a pathogen or malignant cell

10   to grow in the presence of different therapeutic agent(s) in the laboratory. This is usually expressed as the fold-change in the IC_{50} or IC_{90} values (the IC_{50} or IC_{90} value being the therapeutic agent concentration at which 50% or 90% respectively of the population is inhibited from replicating). A highly resistant pathogen or malignant cell might show a 50 or 100-fold increase in IC50, for example.

15   Phenotyping is a direct measure of susceptibility, reflecting the effects and interactions of all the mutations, known or unknown, on the behavior of the pathogen or malignant cell population in the presence of therapeutic agent(s). Any method capable of measuring changes in the ability of a pathogen or malignant cell to grow in the presence of a therapeutic agent(s) can be used in the present invention. Such

20   methods of phenotyping a pathogen or a malignant cell are known to persons of skill in the art.

For example methods for phenotyping bacteria are described in Guoming *et al.*, *Sex. Transm. Dis. 27(2)*:115-8 (2000), Lozano-Chiu *et al.*, *Diagn. Microbiol. Infect. Dis. 31(3)*:417-24 (1998), Iwen *et al.*, *J. Clin. Microbiol. 34(7)*:1779-83 (1996), all

25   expressly incorporated herein by reference.

As an additional illustrative example, methods for phenotyping viruses include, but are not limited to, plaque reduction assays, PBMC p24 growth inhibition assays (see, e.g., Japour *et al.*, *Antimicrob. Agents Chemother. 37*:1095-1101 (1993); Kusumi *et al.*, *J. Virol. 66*:875-885 (1992), recombinant virus assays (see, e.g., Kellam &

30   Larder, *Antimicrob. Agents Chemother. 38*:23-30 (1994); Hertogs *et al.*, *Antimicrob. Agents Chemother. 42*:269-276 (1998), all of which are expressly incorporated herein by reference; the use of GFP as a marker to assess the susceptibility of anti-viral inhibitors (Marschall et al., Institute of Clin. and Mol. Virol., University of Erlanger-Nuremberg, Schlobgarten, Germany); and cell culture assays (Hayden et al., *N. Eng. J.*

35   *Med. 321*:1696-702 (1989), herein incorporated by reference).

As yet another illustrative example, methods for phenotyping malignant cells include, but are not limited to, flow cytometric assays (see, e.g., Pallis *et al., Br. J.*

-16-

Haematol. 104(2):307-12 (1999); Huet *et al.*, *Cytometry 34(6)*:248-56 (1998), both of which are expressly incorporated herein by reference), fluorescence microscopy (see, e.g., Nelson *et al.*, *Cancer Chemother. Pharmacol. 42(4)*:292-9 (1998), expressly incorporated herein by reference), calcein accumulation method (see, e.g., Homolya *et*
5   *al.*, *Br. J. Cancer. 73(7)*:849-55 (1996), expressly herein incorporated by reference), and ATP luminescence assay (see, e.g., Andreotti *et al.*, *Cancer Res. 55(22)*:5276-82 (1995), expressly incorporated herein by reference).

　　　　Under one preferred embodiment, the phenotype methodology employed in the present invention uses a detection enhancer. As used herein, a detection enhancer, or
10   domain, may be a resonant, coloured, colourogenic, immunogenic, fluorescent, luminescent, or radioactive probe. In one embodiment, a detection part encompasses a transcriptional regulator, such as the heterologous reporter system described in U.S. Patent No. 5,776,675, herein incorporated by reference.

　　　　In one embodiment of the invention, the detection enhancer may comprise one
15   or more components of a Fluorescence resonance energy transfer (FRET) system. Such aspects may also be used to design high throughput screening assays. FRET is a process in which an excited fluorophore (a resonance donor) transfers its excited state energy to a light absorbing molecule (a resonance acceptor). Detection enhancers have been successfully used in the phenotyping of HIV-1. Pauwels *et al.*, *J. Virol. Methods*
20   *20*:309-321 (1998); Paulous *et al.*, International Workshop on HIV Drug Resistance, Treatment Strategies and Eradication, St. Petersburg, Florida, USA. Abstr. 46 (1997); and Deeks *et al.*, 2nd International Workshop on HIV Drug Resistance and Treatment Strategies, Lake Maggiore, Italy. Abstr. 53 (1998), all of which are herein incorporated by reference.

25   　　　　Under one preferred embodiment, a phenotype-genotype database is generated to correlate each of the known genotype mutations with changes in the phenotypic drug resistance of that pathogen or malignant cell. By generating such a database, the initial set-up time for the neural network is substantially reduced for the information from such databases are used to train and test the neural networks of the present invention.
30   In certain circumstances, such phenotype-genotype databases have already been generated. It is understood, however, that the present invention can be practiced by establishing a phenotype-genotype database concurrently with the establishment and training of the neural network.

　　　　Under another preferred embodiment of the present invention, a phenotype-
35   genotype database is developed that correlates known genotype mutations with the development of a genetic disease. Preferably, the genotype mutations are indirectly correlated with the development of a genetic disease. Genetic mutations correlated with the development of a genetic disease are generally known to person of skill in the

-17-

art. For example, mutations in the p53 gene are correlated with the development of a number of genetic diseases (Gallagher *et al.*, *Ann. Oncol.* *10*:139-50 (1999); Lenz *et al.*, *Clin. Cancer Res.* *4*:1243-50 (1998); Trepel *et al.*, *Leukemia* *11*:1842-1849 (1997); Iwadate *et al.*, *Int. J. Cancer* *69*:236-40 (1996), all of which are herein incorporated by

5   reference). Likewise, and by way of illustration, many diseases have been linked to genetic mutations, including thyroid diseases (Finke, *Exp. Clin. Endocrinol. Diabetes* *104 Suppl. 4*:92-97 (1996), herein incorporated by reference); Alzheimer disease (Roses, *Neurogenetics* *1*:3-11 (1997), herein incorporated by reference); endometriosis (Bischoff *et al.*, *Hum. Reprod. Update* *6*:37-44 (2000), herein incorporated by

10   reference); hereditary bone tumors (McCormick *et al.*, *Mol. Med. Today* *5*:481-486 (1999), herein incorporated by reference); breast cancer (Chen *et al.*, *J. Cell Physiol.* *181*: 385-92 (1999); Beckmann *et al.*, *J. Mol. Med.* *75*:429-39 (1997), both of which are herein incorporated by reference); and.cervical carcinoma (Lazo, Br. J. Cancer *80*:2008-18 (1999), herein incorporated by reference).

15   It is understood that under one embodiment of the present invention, the new mutation(s) or mutation patterns are added to the phenotype-genotype database. Thus, by use of the present invention, one is able to add to the phenotype-genotype database, thereby further expanding the knowledge and capabilities of the neural networks of the present invention. Furthermore, it is understood that the new mutation(s) or mutation

20   patterns identified through the neural networks of the present invention can be outputted into a report. Such reports can be used by the skilled to practitioner to screen the genetic profile of a patient's to determine the resistance pattern of the disease.

Because of the time and expense associated with phenotypic testing, these assays are generally not suitable for routine clinical screening. Likewise, because of

25   the difficulties in translating genomic information into meaningful data, genotype screening by itself is not suitable for routine clinical screening. The present invention, however, bridges the gap between the more meaningful data obtained from phenotypic testing and the more readily obtainable data obtained from genotypic testing through the use of a neural network.

30   **C. Neural Networks**

Neural networks make neither the assumption of how outputs depend on inputs nor the assumption that inputs are independent. Instead, neural networks offer a very powerful and general framework for representing non-linear mapping from a set of input variables to another set of output variables. Moreover, neural networks represent

35   non-linear functions of many variables in terms of superposition of non-linear functions of single variables. These non-linear functions of single variables are themselves adapted to the data as part of the training process so that the number of such functions

-18-

only needs to grow as the complexity of the problem itself grows, and not simply as the dimensionality grows. It should be appreciated that apart from non-linear functions, also linear functions which concern only an input and an output layer, can be used for the practise of the instant invention. Adding hidden layers to the latter system requires a
5   non-linear function for resolution of the problem.

The neural network of the present invention is characterized by: (1) its pattern of connections between the neurons (called its architecture); and (2) and the knowledge which is represented by weights on the connection. Figure 3 depicts an exemplary framework for a three-layer neural network.

10   **1. Neural Network Architecture**

According to an aspect of the present invention, a neural network is employed to model the relationship between genotype and phenotype for therapeutic agent resistance testing. According to another aspect of the present invention, a neural network is employed to identify mutation(s) or mutation patterns, including insertions
15   and deletions, that confer resistance to a therapeutic agent. Under yet another aspect of the present invention, a neural network is employed to define the genetic basis of therapeutic agent resistance. Under one embodiment of this aspect, a neural network is employed to identify new mutations associated with therapeutic agent resistance.

Preferably, the neural network of the present invention employs a back-
20   propagation and momentum term learning algorithm implemented with supervised multi-layer perception (MLP) architecture. It is understood, however, that other forms of neural networks can be employed in the present invention. For example, adaline networks, adaptive resonance theory networks, bi-directional associative memory networks, back propagation networks, counter propagation networks, Hamming
25   networks, Hopfield networks, Madaline networks, probabilistic neural networks, recirculization networks, spatio-temporal pattern recognition networks, and other types of neural networks can be used to achieve the objects of the present invention.

A neural network consists of a large number of simple processing elements called neurons (also referred to as nodes). The arrangement of neurons into layers and
30   the connection patterns within and between layers is called the network architecture or architecture. Each neuron is connected to other neurons by means of directed communication links with an associated weight. Each neuron has an internal state, called its activation level, which is a function of the inputs it has received. Under one embodiment of the present invention, the activation level is bounded between 0 and 1.
35   Under another embodiment, the activation level is bounded between −1 and 1.

The neural network of the present invention may, for example be a feed-forward network where the signals flow from the input units to the output units in a forward

direction. The feed-forward network of the present invention is a multi-level feed-forward network with one or more hidden layers. Under one embodiment, the neural network of the present invention employs a single hidden layer.

Under one embodiment, the feed-forward network of the present invention is
5   fully connected where every node in each layer of the network is connected to every other node in the adjacent forward layer. However, it is understood that partially connected networks can also be employed in the present invention. Partially connected networks may be employed where too much mutation or polymorphism input data is applied to the network. Alternatively, pruning techniques can be applied. It is
10   understood that in a partially connected network, some of the communication links are missing from the network.

The action of the feed-forward network is determined by two things -- the architecture and the value of the weights. The numbers of input and output nodes are determined by the number of mutations involved and the number of therapeutic agents
15   being considered and so they are considered to be fixed. Initially, the value of the weights and biases are randomized. As training occurs, which is described in more detail below, the weights are adjusted to reduce the error function.

Figure 3 depicts an exemplary framework for a three-layer neural network. The network has I inputs, J hidden units and K output units, and two bias units both of
20   which have an input signal of 1 (i.e., x_0 and z_0). Preferably, the number of inputs, I, is equal to the number of mutations that are known to correlate to phenotypic therapeutic agent resistance for the disease being evaluated. However, under another embodiment, the number of input units, I, is equal to the number of mutations that are known to correlate to phenotypic therapeutic agent resistance for a gene existing in the disease
25   being evaluated. For example, in HIV-1, the input, I, could equal all of the known mutations known to confer therapeutic agent resistance to HIV or it could equal all of the known mutations known to confer therapeutic agent resistance in the protease gene. Under a further sub-embodiment, only a sub-set of known mutations are inputted into the neural network of the present invention.

30   Each hidden layer, J, contains a plurality of hidden nodes. The number of hidden nodes, j, is considered to be a variable that can be adjusted to achieve good performance. In practice, the optimal number of hidden nodes is determined empirically. The means for determining the optimum number of nodes is well known to those of skill in the art and depends on the complexity of the genotype/phenotype
35   information and disease being solved. Like the number of hidden layers, the number of hidden units also affect the complexity of the neural network. The number of hidden units is determined by evaluation the performance of the neural network on the validation and test sets.

The number of output units, K, may be equal to the number of therapeutic agents with known mutations conferring resistance to the therapeutic agents. However, it is understood that the number of output units, K, can be a sub-set of therapeutic agents with known mutations conferring resistance. For example, the number of output
5 units can be restricted to a particular class of therapeutic agents, such as protease inhibitors, etc.

The exemplary three-layer neural network of Figure 3 has two layers of adaptive weights (w_{ji} and w_{jk}), which are the weight of the jth hidden unit associated with input signal x_i, and the weight of the kth output unit associated with the hidden
10 signal z_j, respectively. The values of these weights are optimized during the training step of the neural network, which is described below.

Under the embodiment of the present invention where mutation(s) and/or mutation pattern(s) are identified that confer resistance to a therapeutic agent, it is preferred that the number of inputs be equal to the number of mutations known to be
15 correlated with conferring resistance to that therapeutic agent. The number of outputs is equal to the number of therapeutic agents being studied by the present neural network for mutation identification.

Where the present invention is used to predict the probability of developing a disease, it is preferred that the number of inputs be equal the number of mutations
20 known to be correlated with the development of the genetic disease(s). Under another embodiment, the number of inputs is equal to the number of mutations known to be correlated with the development of a given genetic disease. The number of outputs, preferably, is equal to the number of genetic disease(s) being evaluated by the neural network.

25 Under one embodiment of the present invention, the neural network employs a sigmoid curve as the activation function. The sigmoid curve can be binary (0, 1) or bipolar (-1, 1). Other activation functions that can be employed are linear, hyperbolic tangent, logistic, threshold and Gaussian functions.

2. Neural Network Training

30 Prior to inputting data into an input node, it must be pre-processed. Pre-processing refers to the process of converting molecular data into an input vector capable of being inputted into the neural network. Under one embodiment of the present invention, the mutation pattern x for a given sample is expressed by $x=(x_1,x_2,...,x_n)$, where x_i ($i=1,2,...,n$) has a value 0 or 1, with 1 representing the mutation occurring at position i,
35 and 0 representing no mutation at position i, and n is the number of mutations in the test sample. The output data, likewise, needs to be pre-processed to convert the neural network data into meaningful data. Under one embodiment of the present invention, a

fold resistance of less than or equal to 4 times the cut-off (it may differ from drug to drug) is considered to be "sensitive," greater than 4 cut-off and less than resistant cut-off (it may differ from drug to drug)is considered to be "intermediate," and if the value is greater than 10 the cut-off, it is considered to be "resistant." Biological cut-offs are

5   determined using the technology described in Larder BA & Harrigan PR. AIDS, 2000, 14 (supplement 4): S111, Abstract P327 and poster. For example (drug – cut-off):

Zidovudine   4.0, Nevirapine   8.0, Delavirdine   10.0, Efavirenz   6.0,

Lamivudine   4.5, Didanosine   2.0, Zalcitabine   2.0, Stavudine 1.75,

Abacavir   3.0, Indinavir 3.0, Ritonavir 3.5, Nelfinavir 4.0, Saquinavir   2.5,

10   Amprenavir   2.0, Lopinavir 2.5

Under one embodiment, the neural network of the present invention employs a back-propagation and momentum term (BPM) learning rule. BPM learning rules have been reviewed by, e.g., Chauvin and Rumelhart, Backpropagation: Theory,

15   Architectures and Applications, Lawrence Erlbaum Assoc., Hillsdale, N.J. (1995), expressly incorporated herein by reference. BPM algorithms provide a computationally efficient method for changing the weights in a feed-forward network with different activation functions.

BPM training involves three stages: feed-forward of the input training pattern;

20   calculation and back-propagation of the associated error; and adjustment of the weights. In the feed-forward phase, the weights remain unaltered throughout the network, and the function signals are computed on a neuron-by-neuron basis. In the back-propagation phase, error signals are computed recursively for each neuron starting at the output layer, and passed backward through the network, layer by layer to derive

25   the error of hidden units. Weights are correspondingly adjusted to decrease the difference between the network's output and the target output. After training, the neural network only computes in feed-forward phase.

It is understood that the values of the free parameters (the weights and the biases) can be determined by minimizing the error function. One preferred error

30   function that can be employed in the present invention is the root-mean-square error function, which is the square root of the sum-of-square errors calculated from all patterns across the training file. Other error functions are known to persons of skill in the art.

Under another embodiment, the neural network of the present invention

35   employs a counter-propagation (CP) program. See, e.g., Wu and Shivakumar, *Nucleic Acids Res.* 22:4291-4299 (1994), expressly incorporated herein by reference. A CP program approximates training input vector pairs by adaptively constructing a look-up table. In this manner, a large number of training data points can be compressed to a

-22-

more manageable number of look-up table entries. The accuracy of the approximation is determined by the number of entries in the look-up table.

Under one embodiment of the present invention, BP and CP algorithms are used in combination. It has been reported that a network employing a combination of the two algorithms more accurately predicted phylogentic classifications than a network employing either algorithm alone. See, e.g., Wu and Shivakumar, *Nucleic Acids Res.* 22:4291-4299 (1994), expressly incorporated herein by reference.

In addition to BP training, other training algorithms can be employed in the present invention. For example, the pocket algorithm, delta rule, Hebb rule, Hopfield rule, Windrow-Hoff rule, adaline rule, and Kohonen rule can be used to train the neural network of the present invention.

In order to create a network having the best performance on new data, the simplest approach is to compare the error function of different networks using data that is independent of that used for training. By comparing the different networks, the effect of network parameter modifications can be easily measured.

Neural network parameters are determined by searching for the best performance on the test data set. With these parameters, a concordance rate of greater than 75% between genotype and phenotype can be achieved. In one embodiment a concordance rate of greater than 85% is achieved such as a concordance rate of greater than 90% is achieved. It is understood, however, that concordance rates of greater than 95% can be achieved through the present invention.

Several internal parameters of the network of the present invention can be fine-tuned with the help of experimental results and experience. For example, the learning rate η (the size step of the minimization process) can be optimized. The convergence speed of the neural network is directly related to the learning parameter. Too small of a leaning rate will make the training process slow, whereas too large of a learning rate may produce oscillations between poor solutions. In general, one may employ large steps when the search point is far from the minimum with decreasing step size as the search approaches its minimum. Suitable approaches for selecting the appropriate learning rate are provided by, e.g., Hassoun, Fundamentals of Artificial Neural Networks, MIT Press, Cambridge, MA (1995), expressly incorporated herein by reference. The learning rate η is set between 0 to 1, preferably 0.1 to 0.9. It is understood that the learning rate depends on the genotype-phenotype information being analyzed by the neural network.

Another internal parameter that can be optimized in the present invention is the momentum term α. Momentum allows the network to make reasonably large weight adjustments as long as the corrections are in the same general direction for several patterns, while using a smaller learning rate to prevent a large response to the error

-23-

from any one training pattern. It also reduces the likelihood that the neural network will find weights that represent a local minimum. The momentum term is normally chosen between 0 and 1. Preferably, the momentum α is set to 0.9.

Under one embodiment, a data set of genotypic and phenotypic data is

5   collected. For example, the data set may be collected from a phenotype-genotype database. Under one embodiment of the present invention, each member of the data set corresponds to a genetic mutation that is correlated to a phenotypic change in therapeutic agent resistance. Preferably the data set is divided into a training data set and a testing data set. It is not necessary to have a large training data set. If the

10   samples in the training data set represent all possible cases with adequate statistical significance, the addition of new samples generally does not increase the amount of information in the training samples. Instead, it may decrease the useful amount of information to noise ratio in the samples. On the other hand, too small of a training data set will generally not cover all possible variations in the population. The resultant

15   network often simply memorizes the data in the training data set and does not generalize properly.

During training, each member of the training data set is preferably presented to the neural network one datum at a time. For each member of the training data set, the network uses the preprocessed values to estimate a prediction, which is then compared

20   with the actual resistance of the mutation. If the network's prediction is correct, the connection strengths and thresholds within the network are not changed and the next datum is presented. If the estimate of the prediction is not correct, the connection weights and thresholds in both the hidden layer and the output layer are adjusted to reduce the size of the error function. After the adjustments have been made, the next

25   datum is presented. Training need not continue until the error actually meets its minimum. Training can be stopped once a threshold value for the error function (called tolerance) has been reached, or a fixed upper limit on the number of training iterations (called epochs) has been reached. Where error tolerance is used to determine the end-point of training, it is preferred that the error tolerance γ has a value between 0.1 and

30   0.0001. Under another embodiment, training is stopped once about 10,000 epochs have occurred.

Under one embodiment of the present invention, the training step is performed in an iterative fashion. In other words, a first training data set is selected from a phenotype-genotype database for training. This data set is then used to train the neural

35   network. After the network has been trained, the prediction rate or concordance rate of the network is determined from a test data set. Samples which give an incorrect prediction are removed from the test data set and placed into a second training data set. The second training data set comprises the first data training set plus any samples that

-24-

gave an incorrect prediction from the test data set. The second training data set is then used to re-train the neural network. If necessary, this process can be repeated until the desired performance level is achieved. By re-training the neural network in this fashion, it is possible to increase the performance of the neural network.

5   Occasionally, after the network has been trained and testing has begun, it is determined that the number of input units is excessive. When the number of input units is excessive, network training can be slowed and poor generalization can occur. The determination of what is an excessive number of inputs can be a subjective determination and depends on the specific network. However, if it is determined that 10 the number of input units is excessive, it is preferred to reduce the number of input units. Therefore, under one embodiment, input trimming is used to reduce the dimensionality of the input data.

    Under one embodiment of the present invention, a feature detector is employed that extracts salient features from the input data before presenting it to the neural 15 network. For example, a data partition algorithm can be employed to sort non-spare data out, from which a testing set can be randomly selected. One such data partition algorithm is defined as follows:

$$d = \sum_{i=1}^{n} x_i - z_i |$$

    This algorithm calculates the distance (d) between any two mutation patterns (x and z), and makes it possible to sort spare data and noisy data out and avoid selecting 20 them as testing members. The variable - n - is equal to the number of input units. If the neural network continues to fail to correctly classify large portions of the samples in the training data set, even after repeated adjustments to the training algorithm parameters, the neural network complexity should be increased. On the other hand, if the neural network achieves a high rate of correctly classifying the training set, but fails to 25 accurately classify a large number of samples in the testing data set, the network structure is probably too complex for the problem being solved. If this is the case, the number of nodes in the hidden layer(s) should be gradually reduced or if there are multiple hidden players, the number of hidden layers should be reduced.

    Once the neural network has been trained, the network is ready and capable to 30 predict the resistance of a disease to a therapeutic agent based upon the determined genetic sequence of the disease. To make this prediction, a patient sample containing a sample of the disease is isolated and the genetic information of the disease is determined. This determined genetic information is then pre-processed and loaded into the trained neural network. The trained neural network then computes the predicted

-25-

resistance of the disease to a therapeutic agent. Under another embodiment, the neural networks of the present invention are used to identify additional mutation(s) and/or mutation pattern(s), including insertions and deletions, that confer resistance to a therapeutic agent. In accordance with this embodiment, a first set of genetic mutations

5   is identified. The first set of genetic mutations consists of genetic mutations known to confer therapeutic agent resistance. Such mutations are often known to persons of skill in the art and can be obtained from both the internet and from peer-reviewed journals. For example, the Stanford database <http://hivb.stanford.edu/hiv/>> provides a database of mutations known to confer therapeutic agent resistance to HIV-1.

10   However, to the extent a suitable database is not available, one can be readily generated by a person of skill in the art.

Further identified is a second set genetic mutations that consists of genetic mutations present at relatively high frequency in a disease that is resistant to a therapeutic agent. The second set of genetic mutations contains all genetic mutations in

15   the first set plus at least one additional mutation(s). These additional mutations can be any mutation and/or polymorphism that are related to a disease. Often these mutations are known to be associated with therapeutic agent resistance, but it is not known whether the mutations actually confer resistance. Under one embodiment, the extra mutations are selected from those that are present at relatively high frequency in a

20   disease that is resistant to a therapeutic agent. The threshold level of frequency at which the mutations are found in the disease is set by the user for inclusion in the second set. Generally, the threshold level of frequency can range from 1% to 85%. However, under a preferred embodiment, the threshold level is set between 5 to 50% such as 5 to 25%. Under another embodiment, the second set of genetic mutations

25   consists of the first set of genetic mutations and the next 5 to 100 most frequent mutations associated with resistance. However, any number of additional mutations can be included in the second set so long as the mutation occurs at statistically significant rate.

A first neural network and a second neural network are created wherein the

30   number of inputs for each neural network are equal to the number of genetic mutations being studied. Thus, the number of inputs for the first neural network is equal to the number of mutations in the first set of genetic mutations. The outputs for each neural network are equal to the fold resistance being evaluated. For example, the first and second neural networks are trained using the same training data set.

35   After the two networks are well trained, the neural networks can make a prediction as to the phenotypic impact of a mutation on the resistance of the disease to a therapeutic agent. Under one embodiment, a testing data set is run through the first and second trained neural networks. For each of the neural networks, the output, the

-26-

"virtual" phenotype, for each member of the test data set is compared with the "real" phenotype to determine the difference between the "real" and "virtual" phenotypes. Because each member of the test data set is drawn from a data base, the "real" phenotype (the fold resistance to the therapeutic agent being studied) is known. By comparing the difference between the phenotypic prediction of the first and second trained neural networks, new genetic mutations are identified that are predicted to confer therapeutic agent resistance to the disease. Samples where improved prediction (smaller difference between the "virtual" and "real" phenotype) are observed and called improved samples. All extra mutations contained in these improved samples are screened out. The frequency of a mutation being found in the improved sample is compared with that of the mutation being found in the whole samples evaluated in the neural network. A mutation contained in the improved samples is considered to be conferring resistance to the given therapeutic agent when a much larger difference between the two frequencies is observed. The threshold of frequency difference is specified by the user. Generally, the threshold of difference can range from 1 to 50%, preferably it is at least 5%. Under a preferred embodiment, the threshold difference is at least 9%. Under another preferred embodiment, the threshold level of difference is at least 12%. The user can then perform additional experimentation, such as site directed mutagenesis, to confirm that the mutation does confer phenotypic drug resistance.

The trained neural networks of the present embodiment can also identify that a mutation previously associated with resistance to one therapeutic agent additionally confers resistance to another therapeutic agent.

The outputs of the system are continuous variables which originally provide fold change in IC50, then according to the cut-offs used, the samples may be further classified into sensitive, intermediate or resistant to specific drugs

Where the present invention is used to predict the development of a genetic disease in a patient, the neural network is trained in accordance with these methods using a training data set obtained from a phenotype-genotype database of known mutations that are correlated with the development of a genetic disease. Once the network has been trained, the genetic information from the patient sample is determined. Genetic mutations are identified from this sample and these genetic mutations are inputted into the trained neural network. The trained neural network is then able to make a prediction of the likelihood that these genetic mutations will lead to the development of a genetic disease in the patient.

The following examples are provided by way of illustration and are not intended to be limiting of the present invention.

-27-

EXAMPLE 1: Modeling the relationship between genotype and phenotype for HIV
(Human Immunodeficiency Virus) type 1 drug resistance

A. Genotyping Experiments

HIV-1 RNA was extracted from 200 μl of patient plasma using the QIAamp™
5   viral RNA extraction kit (Qiagen, Santa Clarita, CA), according to the manufacture's
instructions. cDNA encompassing part of the *pol* gene was produced using Expand
RT™. A 2.2kb fragment encoding the protease and reverse transcriptase (RT) regions
was then amplified by nested PCR. This genetic material was subsequently used in
both phenotyping and genotyping experiments. See, e.g., Larder *et al.*, *Antimicrob.*
10 *Agents Chemother. 43(8)*:1961-1967 (1999), expressly incorporated herein by
reference. The PCR products obtained from patient plasma samples were genotyped by
dideoxynucleotide-based sequence analysis, using Big Dye™ terminators (Applied
Biosystems) and resolved on an ABI 377 DNA sequencer. See, e.g., Larder et al.,
Antimicrob. Agents Chemother. 43(8):1961-1967 (1999).

15 ### B. Phenotypic Experiments

Phenotypic susceptibility was determined using a MT-4 cell viral cytopathic
effect protection assay. See, e.g., Kashiwase *et al.*, *Chemotherapy 45(1)*:48-55 (1999),
expressly incorporated herein by reference; Larder et al., *Antimicrob. Agents
Chemother. 43(8)*:1961-1967 (1999). Fold resistance values are derived by dividing the
20 mean 50% inhibitory concentration (IC_{50}) for a patient's recombinant virus by the mean
IC_{50} for wild-type control virus. The procedure is also described in WO 97/27480.

C. Data Pre-Processing

The genotypic and phenotypic data from a total of 172 samples was collected
from a phenotype-genotype database. Each member of the data set corresponds to a
25 genetic mutation that is correlated to a phenotypic change in therapeutic agent
resistance. Among these samples, 20 were selected randomly as the members of the
testing data set, the remaining 152 samples were selected as the members of the
training data set. A total of 90 mutation positions were identified, 30 in the protease
coding region, and 60 in the reverse transcriptase, as shown in Table 1 and Table 2.

30

Table 1							
Mutations in the protease region							
10 I	10 R	10 V	20 M	20 R	24 I	30 N	32 I
33 F	36 I	46 I	46 L	47 V	48 V	50 V	54 A
54 L	54 V	71 T	71 V	73 S	77 I	82 A	82 F
82 S	82 T	84 V	88 D	88 S	90 M		

-28-

Table 2							
Mutations in the reverse transcriptase region							
41 L	44 A	44 D	62 V	65 R	67 N	69 D	69 N
69 S	70 E	70 R	74 I	74 V	75 I	75 M	75 T
77 L	98 G	98 S	100 I	101 E	101 Q	103 N	103 Q
103 R	106 A	106 I	108 I	115 F	116 Y	118 I	151 M
179 D	179 E	181 C	181 I	181 V	184 I	184 V	188 C
188 L	189 I	190 A	190 Q	190 S	208 Y	210 W	211 K
211 Q	214 F	215 C	215 F	215 Y	219 E	219 Q	233 V
236 L	238 T	333 D	333 E				

For a given sample, its mutation pattern x is expressed by $\mathbf{x}=(x_1, x_2, ..., x_{90})$, where x_i ($i = 1, 2, ..., 90$) has a value 0 or 1, with 1 representing the mutation occurring

5   at position i, and 0 representing no mutation at position i.

The output variables y are represented by $\mathbf{y}=(y_1, y_2, ..., y_{15})$, with

y_k ($k = 1, 2, ..., 15$) denoting the fold resistance to drug k. They have values, which may differ by several orders of magnitude. By pre-processing, they were arranged for all of the outputs to be of order unity. For each variable, its maximum y^i_{max} and

10   minimum y^i_{min} with respect to both training and testing data sets was calculated. A set of re-scaled variables is given by:

$$\tilde{y}^n_i = \frac{y^n_i - y^i_{min}}{y^i_{max} - y^i_{min}} * a + b$$

According to this formula, y_i denotes fold resistance to drug i, y^i_{min} denotes the minimum of y_i in the whole samples, y^i_{max} denotes the maximum of y_i in the whole

15   samples, n denotes the index of a specific sample, y^n_i denotes fold resistance of the specific sample before pre-processing, $\tilde{y}^n_i$ denotes fold resistance of the specific sample after pre-processing, [b,a] is an interval to which fold resistance values are normalized, usually taken as [0,1].

D. Neural Network Implementation

20   In this example, a three-layer feed-forward neural network architecture was employed, with full interconnections from input units to hidden units and full interconnections from hidden units to output units. The input nodes were used to represent the genotypic mutations, and the output nodes the degrees of resistance to therapeutic agents, with their values denoting the fold resistance to each therapeutic

25   agent. The hidden nodes were used to determine a suitable model order and achieve

-29-

good performance. A back-propagation momentum algorithm (BP algorithm) was also used. The BP algorithm involves an iterative procedure for minimizing an error function, with back-propagation recursively computes the gradient or change in error with respect to each weight in the network and these values were used to modify the
5   weights between network units.

Three layered neural network estimators, comprising 90 input units, 15 output units, and a single hidden layer with the number of units varying from 8 to 26, were trained and tested. The learning rate η was set to 0.1-0.9, the momentum α was set to 0.9, and the error tolerance γ , 0.1-0.0001. Training was terminated when the error
10   tolerance was attained or when 10,000 epochs occurred, whichever happened sooner.

The training and testing results demonstrate that the neural network estimators with inadequate hidden units gave poor predictions for new data, and those with too many hidden units also exhibit poor generalization as shown in Figure 4(a). The results also demonstrate that the performance did not get better when the error tolerance
15   decreased, as shown in Figure 4(b). In Figure 4(b), the error tolerance index was 0.1 for Index 1, 0.05 for Index 2, 0.01 for Index 3, 0.005 for Index 4, 0.001 for Index 5, 0.0005 for Index 6 and 0.0001 for Index 7. This means good generalization was achieved by stopping training at an early stage. In these three-layered neural network estimators, the relevant network parameters were h (the number of hidden units), η,
20   α, and γ . A search in the parameter space showed that the optimal values of h, η, α, and γ are 12, 0.45, 0.9, and 0.01, respectively. With these neural network parameters, the performance of the neural network was evaluated based on 20 testing samples (each with 15 drugs), which were selected randomly from the same database as the training samples. Results from the test samples are summarized in Table 3.
25

Table 3 : Drug resistance level and its prediction

	1	2	3	4	5	6	7	8	9	10
AZT	S/S	I/R	R/R	S/R	S/I	R/S	S/S	S/R	I/S	I/I
3TC	S/S	R/R	R/R	S/S	R/R	R/S	R/R	R/R	R/R	S/S
DDI	S/S	S/S	S/S	S/R	I/I	I/S	S/I	S/S	S/S	S/S
DDC	S/S	S/S	S/S	S/S	S/S	I/S	S/S	S/I	S/S	S/S
D4T	S/S	I/I	I/S	S/S	S/S	I/S	I/I	S/I	S/S	S/S
1592 U89	S/S	S/I	S/S	S/S	S/S	I/S	S/S	I/S	S/S	S/S
PMEA	S/S	S/S	S/S	S/S	S/S	S/S	R/R	S/S	S/S	S/S
Nevirapine	S/S	S/S	R/S	R/R	S/S	R/R	R/R	R/I	S/I	R/R
Delavirdine	S/S	S/S	R/R	R/R	S/S	R/I	R/R	R/R	S/S	R/R
DMP266	S/S	S/S	R/R	R/R	S/S	R/I	S/S	R/R	S/S	R/I

-30-

	1	2	3	4	5	6	7	8	9	10
Indinavir	S/S	S/S	I/R	I/R	R/R	R/S	S/S	I/R	S/S	S/R
Ritonavir	S/S	R/R	R/R	I/R	R/R	R/R	S/S	R/R	S/S	S/R
Nelfinavir	S/S	R/S	I/R	R/R	R/R	R/R	S/S	R/R	R/S	R/R
Saquinavir	S/S	S/S	I/R	I/I	I/I	I/S	S/S	S/R	S/S	S/I
VX-478	S/S	S/I	S/S	S/I	S/I	S/S	S/S	S/S	S/S	S/S

In Table 3, R stands for resistance, S, for sensitive, and I, for intermediate. A fold resistance of less than or equal to 4 is considered to be "sensitive," greater than 4 and less than 10 is considered to be "intermediate," and if the value is greater than 10, it is considered to be "resistant." The symbol "R/I" in Table 4 means that a sample is resistant to a drug from the phenotypic data and was predicted to be intermediate by the neural network model.

AZT (3'-azido-3'-deoxythymidine), ddI (2',3'-dideoxyinosine), PMEA (also known as adefovir, and 9-(2-phosphonylmethoxyethyl)adenine), VX-478 (also known as Amprenavir, Agenerase, and 141-W94) are approved potent inhibitors of a number of viruses.

Simulation experiments were also conducted by combining different mutation patterns. A total of 12 mutation patterns, as shown in Table 4, were added to the testing data sets. The magnitudes of resistance that were simulated are shown in Table 4, and plotted in Figure 5. It can be seen from the simulation results that the development of a 184V mutation can re-sensitize AZT-resistant virus if the 41L and 215Y mutations are already present in the RT of HIV-1. This confirms the biological observation that recombinant viruses containing the 184V mutation in the background of AZT resistance, such as 41L, 67N, 70R, 215Yand 219Q, cause a suppressive effect that result in reversion to AZT sensitivity. The results also demonstrate that the 184V mutation has a strong effect on conferring 3TC resistance no matter what other mutations are involved.

Table 4 : Simulating fold resistance conferred by mutation(s)

Index	Mutation(s)	Fold resistance				
		AZT	3TC	Nevirapine	Delavirdine	DMP266
P1	103N	9.4	5.2	74.8	115.8	238.0
P2	184V	0.5	68.9	0.9	0.7	0.7
P3	77I, 184V	0.7	74.8	2.6	2.9	3.0
P4	103N, 184V	1.4	39.9	30.7	102.6	168.0
P5	41L	8.8	2.3	0.3	0.2	1.6

| Index | Mutation(s) | Fold resistance | | | | |
		AZT	3TC	Nevirapine	Delavirdine	DMP266
P6	215Y	13.7	1.3	0.2	0.1	0.3
P7	41L, 184V, 215Y	2.1	50.9	0.4	0.4	0.5
P8	67N, 219Q	22.7	2.8	0.2	0.1	0.6
P9	67N, 184V, 219Q	5.1	61.5	0.2	0.2	4.9
P10	67N, 70R, 184V, 219Q	4.1	81.0	0.2	0.1	3.3
P11	67N, 70R, 215Y	22.5	3.0	0.2	0.1	0.4
P12	67N, 70R, 215Y, 219Q	41.5	4.8	0.2	0.1	0.3

EXAMPLE 2: Predicting HIV-1 Protease Inhibitor (PI) phenotypic resistance from PI genotype

In this example, the genotypic and phenotypic data from 1162 HIV-1 PI samples was collected from a genotype-phenotype database. A PI genotype refers to a genotype with a mutation or polymorphism in the protease coding region which is considered to conger resistance to a protease inhibitor. A total of 30 mutations were identified in the protease coding region, as shown in Table 1. For a given sample, its mutation pattern x was expressed by $x = (x_1, x_2, ..., x_{30})$, where x_i ($i = 1, 2, ..., 30$) has a value 0 or 1, with 1 representing the mutation occurring at position i, and 0 representing no mutation at position i.

Table 5 : Drug resistance level and its prediction

	IDV	RTV	NFV	SQV	APV
1	S/S	S/S	S/S	S/S	S/S
2	S/S	S/S	R/R	S/S	S/S
3	S/S	S/S	R/R	S/S	S/S
4	R/R	R/R	R/R	R/R	I/I
5	R/I	R/R	R/R	R/R	S/S
6	S/S	S/S	S/I	S/S	S/S
7	S/S	S/S	R/R	S/S	S/S
8	S/S	S/S	R/R	S/S	S/S
9	S/S	S/S	R/R	S/S	S/S
10	S/S	S/I	R/I	S/S	S/S
11	S/I	S/R	I/I	S/I	S/S
12	S/S	S/S	R/S	S/S	S/S
13	R/R	R/R	R/R	R/R	I/I

-32-

	IDV	RTV	NFV	SQV	APV
14	R/R	R/R	R/R	R/R	S/S
15	S/S	S/S	R/R	S/S	S/S
16	R/R	R/R	R/R	R/R	R/R
17	R/R	R/R	R/R	R/R	S/R
18	S/S	S/S	R/R	S/S	S/S
19	S/S	S/S	R/R	S/S	S/S
20	S/S	S/S	R/R	S/S	S/S
21	S/R	R/R	R/R	R/R	S/S
22	S/S	I/I	R/I	I/S	S/S
23	R/R	R/R	R/R	R/R	R/I
24	R/R	R/R	R/R	R/R	R/R
25	I/R	R/R	R/R	R/R	S/I
26	R/R	R/R	R/R	I/I	S/S

Distance d between mutation pattern x and mutation pattern z was defined as follows:

$$d = \sum_{i=1}^{30} | x_i - z_i |$$

By calculating distance between any two mutation patterns, the distribution of the samples in a space was estimated. This made it possible to sort spare data and noisy data out and avoid selecting them as testing members.

Three layered neural network estimators, comprising 30 input units, 5 output units (corresponding to 5 PI drugs) and a single hidden layer with the number of units varying were trained and tested. The performance of neural networks, which were trained with or without noisy data involved in the training set, is shown in Figure 6 and Figure 7. Concordance rates were from 76% for amprenavir (APV) to 93% for ritonavir (RTV) with an average of 82% for the network trained with noisy data. Concordance rates without noisy data were from 79% for amprenavir to 91% for nelfinavir (NFV) with an average of 86%. Thus, better performance was achieved when the noisy data was taken out from both training set and testing set.

Table 6 : Drug resistance level and its prediction.

	IDV	RTV	NFV	SQV	APV
27	R/R	R/R	R/R	S/R	S/S
28	R/R	R/R	R/R	S/S	S/S
29	R/R	R/R	R/R	R/R	S/S
30	R/R	R/R	R/R	R/R	I/I

-33-

	IDV	RTV	NFV	SQV	APV
31	R/R	R/R	R/R	R/R	R/I
32	R/R	R/R	R/R	R/R	R/I
33	I/R	R/R	R/R	I/I	S/S
34	S/S	S/I	R/R	S/S	S/S
35	R/R	R/R	R/R	I/R	I/S
36	R/R	R/R	R/R	R/R	I/I
37	I/I	S/R	R/R	S/S	S/S
38	R/R	R/R	R/R	R/R	I/S
39	R/R	R/R	R/R	R/R	I/I
40	R/R	R/R	R/R	R/R	S/I
41	I/S	I/I	R/R	I/S	S/S
42	S/S	I/S	I/I	R/S	S/S
43	I/S	R/R	I/S	S/S	S/S
44	I/R	R/R	R/R	R/R	I/I
45	R/R	R/R	R/R	R/R	S/S
46	R/R	R/R	R/R	R/R	I/I
47	R/R	R/R	R/R	R/R	I/S
48	R/R	R/R	R/R	R/R	S/S
49	R/R	R/R	R/R	R/R	I/I
50	S/S	R/I	S/S	S/S	S/S
51	R/R	R/R	R/R	R/R	S/I
52	S/S	S/S	R/R	S/S	S/S
53	R/R	R/R	R/R	R/R	I/S

For the later neural network, the testing results are shown in Tables 5 and 6.
With the same network, simulation experiments were conducted by combining different
mutation patterns. The magnitudes of resistance that were simulated are shown in
Table 7. The simulation results demonstrate that nelfinavir exhibits resistance with
even a single mutation 30N or double mutations involved. This makes it different than
other PI inhibitors. Resistance to indinavir (IDV), ritonavir and saquinavir (SQV)
involve multiple mutations, usually greater than three mutations while resistance to
amprenavir requires at least four mutations.

Table 7 : Magnitude of resistance inferred from the model

Mutation(s)	Fold resistance				
	IND	RTV	NFV	SQV	APV
10I	1.1	3.2	1.4	0.2	0.1

-34-

Mutation(s)	Fold resistance				
	IND	RTV	NFV	SQV	APV
30N	1.1	2.0	13.9	0.7	0.5
36I	1.9	3.0	5.4	0.6	0.2
46I	1.2	3.4	2.7	0.2	0.1
71I	1.4	2.0	3.6	0.4	0.2
73S	2.4	5.0	6.8	0.7	0.1
82A	0.4	1.0	0.5	0.1	0.1
84V	3.8	8.3	8.4	4.0	1.0
88D	1.1	3.8	1.3	0.3	0.1
90M	1.5	6.2	4.3	1.3	0.1
30N 77I	1.3	0.2	20.0	0.9	0.3
77I 88S	2.3	1.8	13.0	2.1	0.8
36I 84V 90M	22.8	39.0	37.0	30.8	9.2
54V 71V 73S	17.9	51.1	44.8	5.0	0.4
82A 84V 90M	10.6	34.8	10.3	12.6	4.7
48V 84V 90M	12.4	21.9	20.8	20.6	6.0
10I 46I 84V 90M	34.6	68.9	52.5	31.6	14.9
36I 46I 71V 84V	33.2	74.0	47.0	13.5	11.1
46I 77I 84V 90M	5.9	77.6	54.9	35.2	12.9
10I 46I 71V 84V 90M	17.9	42.9	24.4	14.6	10.9
10I 46I 71V 77I 84V 90M	45.0	77.3	58.4	34.0	11.9
10I 54V 71V 73S 77I 84V 90M	34.3	108.5	69.0	49.1	10.7
10I 33F 71V 77I 84V 88D 90M	9.3	43.8	12.0	16.5	9.1
10V 20M 36I 54V 71V 82A 84V 90M	26.7	186.8	41.6	44.4	9.7

EXAMPLE 3: The application of neural networks in predicting phenotypic resistance from genotypes for HIV-1 protease inhibitors

In this example, a three-layer neural network model was constructed with 30 input nodes, corresponding to 30 mutations in the protease coding region and 5 output nodes, representing the fold resistance values for 5 protease inhibitors. A total of 1068 samples were selected from an HIV-1 phenotype-genotype database. Among these samples, 210 were selected as the testing data set, the remaining samples as the training data set. The performance of the neural network models was evaluated by calculating the prediction rate (concordance rate) in the test data set. An average prediction rate of 76% for 5 protease inhibitors was achieved for these data sets. In order to improve this prediction rate, samples that gave an incorrect prediction were removed from the test

-35-

data set to the training data set and the neural network models was re-trained (with a training data set of 1015 samples and a test data set of 53 samples). With the re-trained neural network, an average prediction rate of 87% in the new test data set and an average concordance rate of 88% in the whole data set were obtained.

5   Next, an additional 60 protease gene polymorphisms were added to the input layer of the neural network model using the same training and test data sets. After training, the neural network gave an average prediction rate of 91% using the same new test data set and an average concordance rate of 92% in the whole data set. Linear regression analysis of the predicted versus actual fold resistance gave an r^2 value of

10   0.85 for the test data set. Analysis of this data set indicates that the improvement in prediction was due to the additional polymorphisms added to the model, such as 13V, 55R, 57K and 93L.

EXAMPLE 4:   Modeling the relationship between genotype and phenotype for stavudine (d4T) using neural networks

15   In this example, a total of 1182 samples with >4 fold d4T resistance were selected from a phenotype-genotype database for analysis. 105 samples were selected randomly as a test data set, the remainder was used as a training data set. By searching for the most frequent RT mutations in the database that are associated with stavudine resistance, 57 RT mutations were identified and used as the input variables for the

20   neural network models. Following training, a prediction rate of 72% in the test set was achieved. In order to improve this prediction rate, samples which gave an incorrect prediction were removed from the test data set into the training data set and the neural network models were re-trained with a training data set of 1041 samples and a test data set of 41 samples. As a result, an average prediction rate of 85% in this new test set

25   was achieved. Among these predictions, 84% gave the correct prediction of intermediate/intermediate (>4 fold change <10 fold change in stavudine sensitivity) and 89% gave the correct prediction of resistant/resistant (>10 fold resistance). 16% of the samples gave the incorrect intermediate/resistant prediction and 11% gave a resistant/ intermediate prediction. Linear regression analysis of the predicted versus actual fold

30   resistance gave a r^2 value of 0.67 for the test data set. These results demonstrate that the performance of the neural network model can be improved as the size of training data set is increased.

The neural network prepared according to this example was also able to identify mutation patterns that confer resistance to stavudine. Mutations previously known to

35   confer stavudine resistance, such as 151M and the "69 insertion" family were highlighted by this analysis. Additional mutational patterns that included AZT resistance mutations were also identified by the neural network as conferring resistance

-36-

to stavudine. From these results, it appears that pathways other than multi-nucleoside resistance can confer stavudine resistance.

EXAMPLE 5:   Another application of neural networks in predicting phenotypic resistance from genotypes for HIV-1 protease inhibitors

5   In this example, the interpretation of HIV-1 drug resistance mutation patterns has been improved by predicting the phenotype using a large phenotype-genotype database. To predict the phenotype from a genotype, the database is searched and phenotypes of samples matching the genotype are retrieved. The "virtual phenotype" is obtained by calculating the average increase in fold resistance for each drug in the

10   matching group. To determine new mutation patterns, neural network techniques were adopted to determine the relationship between genotypes and phenotypes for the 5 HIV-1 protease inhibitors. Three-layer neural network models were constructed with 30 input nodes, corresponding to 30 mutations in the protease coding region and 5 output nodes, representing the fold resistance values for 5 protease inhibitors. A total

15   of 1068 samples were selected from a phenotype-genotype database for HIV-1. Among these samples, 210 were selected as the test data set, the remaining samples as the training data set. The performance of the neural network models was evaluated by calculating the prediction rate in the test data set. An average prediction rate of 76% to 5 protease inhibitors was achieved for these data sets. In order to improve this

20   prediction rate, samples that gave an incorrect prediction were removed from the test data set to the training data set and the neural network models were re-trained (with a training data set of 1015 samples and a test data set of 53 samples). Now, an average prediction rate of 87% in the new test data set and an average concordance rate of 88% in the whole data set were obtained. Next, an additional 60 protease gene

25   polymorphisms were added to the input layer of the neural network model using the same training and test data sets. After training, the neural network gave an average prediction rate of 92% using the same new test data set and an average concordance rate of 93% in the whole data set. Linear regression analysis of the predicted versus actual fold resistance gave an r^2 value of 0.85 for the test data set. Analysis of this data

30   set indicated that the improvement (significant with p=0.036) in prediction was due to the additional polymorphisms added to the model, such as 13V, 55R, 57K and 93L.

          In this Example, a generic framework of modeling the relationship between genotype and phenotype for HIV-1 drug resistance has been developed. Neural network models with 30 identified mutations and 90 mutations/polymorphisms were

35   trained and tested. Improvement of prediction rate was observed and the corresponding additional polymorphisms that lead to the improvement were sorted out. Prediction comparisons were done in both testing data set and the whole data set studied.

-37-

Analysis of this data set indicated that the improvement in prediction was due to the additional polymorphisms added to the model, such as 13V, 55R, 57K and 93L.

A.   Neural network model

A generic framework was developed for modeling the relationship between genotypes and phenotypes of HIV-1 drug resistance as shown in Figure 1. It consists of the following phases: determining NN architecture, collecting data, selecting mutations/polymorphisms and drugs, partitioning data, NN training and test, statistical analysis.

Neural network architecture:

The first step is to design a specific network architecture, including a specific number of "layers" each consisting of a certain number of "neurons." The size and structure of a neural network needs to match the nature of the HIV-1 drug resistance. However, the nature is obviously not known very well at this early stage. In order to determine a suitable network architecture, various networks, with a fixed number of hidden layer and different number of hidden units, were trained using a training data set. The performance of the neural networks was then evaluated and compared using a test set. The neural network architecture was finally determined by selecting the network having the best performance with respect to the test set.

Gathering data for neural networks:

Neural networks learn from existing data. In order to investigate the relationship between genotypes and phenotypes of HIV-1 drug resistance using a NN, data needs to be gathered for training and test purposes. Both genotypes and phenotypes of samples were exported from a database into Excel files. Programs were designed to extract these genotypic data and phenotypic data for each individual sample. Phenotypic data consists of fold resistance to all drugs tested. Genotypic data contains all the polymorphisms in gag, reverse transcriptase, and protease coding regions.

Input and output variables:

The training set and test set include a number of cases, each containing values for a range of input and output variables. The choice of output variables is straightforward, depending on how many and which drugs are considered in the neural network models. The easiest way to select input variables is to consider all polymorphisms, even all sequence strings, as input variables. However, this may lead to a problem what is known as "the curse of dimensionality." As the number of input variables increases, the number of cases required increases non-linearly. In this Example, determining the input variables was guided initially by intuition. Expertise in HIV-1 drug resistance provided some idea of which variables are likely to be

-38-

influential. For example, it is reasonable to select identified mutations and higher frequency polymorphisms as input variables.

Data selection and partitioning:

Selecting data and determining the number of cases required for neural network
5   training presented difficulties. Neural network technologies rely on a key assumption that the training and test data must be representative of the underlying system. A neural network can only learn from cases that are present. If cases of sensitive phenotypes were not included in the training set, it is not expected that the neural network will make a correct decision when it encounters genotypes that associate to
10   sensitive phenotypes. That is to say, the types of cases that are expected to predict must be covered in the training set. Since a neural network minimizes an overall error, the proportion of types of data in the set is also critical. A network trained on an unbalanced data set will bias its decision towards higher proportion of types. If the representation of the proportion of types is different in the real population, the network
15   may not give a good decision. Generally speaking, the best approach for data selection is to ensure even representation of different cases, and to interpret the network's decisions accordingly. In this Example, 1162 cases were selected from the database, with each case having >10 fold resistance to at least one of the drugs. Data analysis showed that conflicting cases existed in the samples exported. These cases made it
20   difficult to improve the performance of neural networks, and were then removed from the samples. Data analysis also demonstrated that data is not evenly distributed in the samples. Compared with the higher dimensional issue, the size of the training set seems still small. In this case, it is not suitable if the sparse cases are selected as test set. To address this issue, a data partition algorithm was designed to sort non-sparse
25   data out, from which a test set was randomly selected. The remainder cases were taken as a training set.

Statistical analysis:

In order to reasonably interpret results, statistical analyses were applied to the evaluation of the correlation between the predicted phenotypes and the actual
30   phenotypes, and the testing of various statistical significances. The correlation coefficient that is far from zero provided four possible explanations about the relationship between the predicted and the actual phenotypes. The conclusion may be: that the predicted phenotypes help determine the values of the actual phenotypes; that another variable may also influence the actual phenotypes besides the predicted
35   phenotypes; that the predicted phenotypes and the actual phenotypes do not correlate at all; or that a strong correlation was observed, as in this case. The p-value determines how often this could occur. The p-value of a result is the probability that the observed relationship in a sample occurred by pure chance, and that in the population from which

-39-

the sample was drawn, no such relationship exists. The r squared provides information about how much percentage of variance is shared between the predicted and the actual phenotypes.

5   **Table 8 : Predicting phenotypes against actual phenotypes**

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
1	S/S	S/S	S/S	S/S	S/S
2	S/S	S/S	R/R	S/S	S/S
3	S/I	S/S	R/R	S/S	S/S
4	R/R	R/R	R/R	R/R	I/I
5	R/I	R/R	R/I	R/I	S/S
6	S/S	S/S	S/S	S/S	S/S
7	S/S	S/S	R/R	S/S	S/S
8	S/S	S/S	R/R	S/S	S/S
9	S/S	S/S	R/R	S/S	S/S
10	S/S	S/S	R/S	S/S	S/S
11	S/I	S/I	I/I	S/S	S/S
12	S/S	S/S	R/S	S/S	S/S
13	R/R	R/R	R/R	R/R	I/I
14	R/R	R/R	R/R	R/R	S/S
15	S/S	S/S	R/R	S/S	S/S
16	R/R	R/R	R/R	R/R	R/R
17	R/R	R/R	R/R	R/R	S/R
18	S/S	S/S	R/R	S/S	S/S
19	S/S	S/S	R/I	S/S	S/S
20	S/S	S/S	R/R	S/S	S/S
21	S/R	R/R	R/R	R/R	S/S
22	S/I	I/I	R/I	I/I	S/S
23	R/R	R/R	R/R	R/R	R/R
24	R/R	R/R	R/R	R/R	R/R
25	I/R	R/R	R/R	R/R	S/S
26	R/R	R/R	R/R	I/S	S/S

Predicting results on the test data set are summarized in tables 8 and 9, where R stands for resistance, S, for sensitive, and I, for intermediate. A fold resistance of less than or equal to 4 is considered to be 'sensitive', greater than 4 and less than 10 is considered

10   to be 'intermediate', and if the value is greater than 10, it is considered to be 'resistant'.

-40-

It should be understood that any cut-off value can be defined for instance biological cut-off as described in Larder BA & Harrigan PR. AIDS, 2000, 14 (supplement 4): S111, Abstract P327

The symbol 'R/T' in tables 8 and 9 means that a sample is resistant to a drug from the phenotypic data and is predicted to be intermediate by the model.

Table 9 : Predicting phenotypes against actual phenotypes

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
27	R/R	R/R	R/R	S/I	S/I
28	R/R	R/R	R/I	S/S	S/S
29	R/R	R/R	R/R	R/R	S/S
30	R/R	R/R	R/R	R/R	I/I
31	R/R	R/R	R/R	R/I	R/R
32	R/R	R/R	R/R	R/R	R/I
33	I/R	R/R	R/R	I/I	S/S
34	S/R	S/R	R/R	S/S	S/S
35	R/R	R/R	R/R	I/I	I/S
36	R/R	R/R	R/R	R/R	I/I
37	I/I	S/I	R/R	S/S	S/S
38	R/R	R/R	R/R	R/R	I/I
39	R/R	R/R	R/R	R/R	I/I
40	R/R	R/R	R/R	R/R	S/I
41	I/I	I/I	R/R	I/S	S/S
42	S/S	I/I	I/I	R/I	S/S
43	I/I	R/R	I/R	S/S	S/S
44	I/R	R/R	R/R	R/R	I/I
45	R/R	R/R	R/R	R/R	S/S
46	R/R	R/R	R/R	R/R	I/I
47	R/I	R/R	R/R	R/R	I/I
48	R/R	R/R	R/R	R/R	S/I
49	R/R	R/R	R/R	R/R	I/I
50	S/S	R/R	S/S	S/S	S/S
51	R/R	R/R	R/R	R/R	S/I
52	S/S	S/S	R/R	S/S	S/S
53	R/R	R/R	R/R	R/R	I/I

An average prediction rate of 87% to 5 protease inhibitors was obtained in the new test data set. When an additional 60 protease gene polymorphisms, as shown in

Table 10, were added to the input layer of the neural network model, the re-trained neural network model gave an average prediction rate of 92% in the same test data set. The predicting results using 90 mutations/polymorphisms are summarized in Tables 11 and 12.

Table 10 : 60 polymorphisms in the protease coding region

8D	8Q	10F	13V	20I	20L	20T	20V
22V	23I	24F	32A	33I	33M	33V	33X
36L	36Q	36R	36T	36V	48T	54S	54T
55R	55T	57K	58E	63A	63C	63H	63I
63N	63P	63Q	63R	63S	63T	63V	71D
71I	71L	73A	73C	73T	82C	82I	82M
84A	84C	84L	85V	88I	88T	89I	89M
89T	89V	93L	93M				

By comparing Tables 8, 9 and Tables 11, 12, it was found that the improvement of phenotype prediction in the test data set happened in 23 of 53 samples, as listed in table 13, where the first letter denotes for the actual phenotype, the second, the predicted phenotype using 30 mutations, and the third, the predicted phenotype using 90 mutations/polymorphisms. The corresponding genotypic differences are summarized in Table 14.

Regression analyses of predicted phenotypes and the actual phenotypes are shown in Figures 8 and 9. In order to test whether the improvement is significant or not, the predicted distributions on the test data set in both cases are summarized in Tables 15 and 16 and the p-values are calculated as follows, S/S (0.187), I/I (0.382), and R/R (0.036). It can be seen that the improvement of predicted phenotypes from R to R is significant, although there is no evidence of significant improvement from S to S and I to I. Similar analyses were also done in the whole samples used in this work. The results indicated that the additional polymorphisms added to the model, such as 13V, 55R, 57K, and 93L, lead to the improvement in prediction.

Table 11 : Predicting phenotypes using 90 mutations/polymorphisms against actual phenotypes

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
1	S/S	S/S	S/S	S/S	S/S
2	S/S	S/S	R/R	S/S	S/S
3	S/S	S/S	R/I	S/S	S/S

-42-

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
4	R/R	R/R	R/R	R/R	I/I
5	R/I	R/R	R/R	R/R	S/S
6	S/S	S/S	S/S	S/S	S/S
7	S/S	S/S	R/R	S/S	S/S
8	S/S	S/S	R/R	S/S	S/S
9	S/S	S/S	R/R	S/S	S/S
10	S/S	S/S	R/S	S/S	S/S
11	S/S	S/S	I/I	S/S	S/S
12	S/S	S/S	R/R	S/S	S/S
13	R/R	R/R	R/R	R/R	I/I
14	R/R	R/R	R/R	R/R	S/S
15	S/S	S/S	R/R	S/S	S/S
16	R/R	R/R	R/R	R/R	R/R
17	R/R	R/R	R/R	R/R	S/I
18	S/S	S/S	R/R	S/S	S/S
19	S/S	S/S	R/R	S/S	S/S
20	S/S	S/S	R/R	S/S	S/S
21	S/R	R/I	R/R	R/R	S/S
22	S/I	I/R	R/R	I/I	S/S
23	R/R	R/R	R/R	R/R	R/R
24	R/R	R/R	R/R	R/R	R/R
25	I/I	R/R	R/R	R/R	S/S
26	R/R	R/R	R/R	I/S	S/S

Table 12 : Predicting phenotypes using 90 mutations/polymorphisms against actual phenotypes

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
27	R/R	R/R	R/R	S/R	S/I
28	R/R	R/R	R/R	S/S	S/S
29	R/R	R/R	R/R	R/R	S/S
30	R/R	R/R	R/R	R/R	I/I
31	R/R	R/R	R/R	R/R	R/R
32	R/R	R/R	R/R	R/R	R/R
33	I/I	R/R	R/R	I/I	S/S
34	S/S	S/I	R/I	S/S	S/S
35	R/R	R/R	R/R	I/I	I/S

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
36	R/R	R/R	R/R	R/R	I/S
37	I/I	S/I	R/R	S/S	S/S
38	R/R	R/R	R/R	R/R	I/I
39	R/R	R/R	R/R	R/R	I/I
40	R/R	R/R	R/R	R/R	S/S
41	I/S	I/R	R/R	I/S	S/S
42	S/I	I/I	I/I	R/R	S/S
43	I/I	R/R	I/I	S/S	S/S
44	I/I	R/R	R/R	R/R	I/I
45	R/R	R/R	R/R	R/R	S/S
46	R/R	R/R	R/R.	R/R	I/I
47	R/R	R/R	R/R	R/R	I/R
48	R/R	R/R	R/R	R/R	S/S
49	R/R	R/R	R/R	R/R	I/I
50	S/S	R/R	S/S	S/S	S/S
51	R/R	R/R	R/R	R/R	S/I
52	S/S	S/S	R/R	S/S	S/S
53	R/R	R/R	R/R	R/R	I/I

Table 13 : Improvement of predicting phenotypes from using 30 mutations to using 90 polymorphisms

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
1			R/I/R	R/I/R	
2	S/I/S	S/I/S			
3			R/S/R		
4					S/R/I
5			R/I/R		
6		R/R/I			
7		I/I/R	R/I/R		
8	I/R/I				
9				S/I/R	
10			R/I/R		
11				R/I/R	
12					R/I/R
13	I/R/I				
14	S/R/S	S/R/I			

-44-

	Indinavir	Ritonavir	Nelfinavir	Saquinavir	Amprenavir
15					I/I/S
16					S/I/S
17	I/I/S	I/I/R			
18	S/S/I		R/I/R		
19				I/R/I	
20	S/I/S		R/R/I		
21	I/R/I				
22	R/I/R				I/I/R
23					S/I/S

Table 14 :Genotypic differences, which lead to improvement of predicting phenotypes

	Identified mutations	Additional polymorphisms
1	10I,71V,73S,84V,90M	63P,85V,93L
2	10I,46I,48V,77I,82A,90M	10E,58E,63T
3	10I	36V,93L
4	10I,46I,71V,77I,84V,90M	63P,93L
5	10V,30N	13V,63P
6	10I,48V,54V,82A	13V
7	71V,73S,90M	20I,63P
8	10I,54V,71V,73S,84V,90M	33M,63P
9	10I,46L,54V,71V,82A,90M	63P,93L
10	46I,71V,73S,90M	20I,63P,93L
11	10I,46L,54V,71V,77I,82A,90M	55R,58E,63P,93L
12	10I,36I,46I,84V,90M	20I,63P,73C,85V
13	10I,36I,71T,90M	63P,73T
14	10I,46I,71T,77I,90M	57K,63P,93L
15	46I,84V,90M	20I,63Q
16	10I,46I,77I,84V,90M	63P,73T,93L
17	46I,77I,90M	10F,20L,63P
18	36I,71T,90M	63P,93L
19	54V,71V,82A	63P,93L
20	77I,88S	13V, 63P, 93L
21	10I,77I,84V,90M	63Q
22	10I,20R,36I,71V,73S,84V,90M	13V,63P
23	10I,46I,77I,90M	20I,63P,73T

-45-

Table 15 : Predicted drug resistance level against the actual ones using 30 mutations

Actual	S	I	R
S	83	10	4
I	3	22	4
R	2	10	127

Table 16 : Predicted drug resistance level against the actual ones using 90 mutations/polymorphisms

Actual	S	I	R
S	88	7	2
I	5	21	3
R	1	3	135

The improvement in prediction by adding new polymorphisms indicated that the NN model has an ability to identify new mutations. Statistical analysis demonstrated that the predicted phenotypes correlate to the actual phenotypes and the results in this example also demonstrated the accuracy of NNs in predicting the magnitude of resistance to protease inhibitors based on genotypic mutations. The performance of the neural network model is expected to improve given that the size of the training samples used was rather small and since an NN becomes more 'knowledgeable' as the number of training samples increases.

EXAMPLE 6: Use of neural networks to define the genetic basis of HIV-1 resistance to d4T.

This Example describes a systematic method that was used to investigate the relationship between mutation patterns and corresponding phenotypic resistance using neural networks. First, a therapeutic agent was selected for study, in this case d4T. Three neural network models (the 9RT, 26RT and 60RT models) were developed to investigate how mutation patterns influence d4T resistance. The 9 RT model was based on the nine mutations listed in the Stanford sequence database (http://www.hivb.stanford.edu) associated with d4T resistance (62V, 69D, 69N, 69SXX, 75I, 75T, 77L, 116Y, and 151M). The other models were based on adding either the next 17 or 51 most frequent RT mutations present in d4T resistant samples. Thus, the 26 RT mutation model included the 9 RT mutation model plus the 17 most frequent mutations in d4T resistant samples. These 17 mutations were 41L, 44D, 67N, 70R, 75A, 75M, 115F,118I,184V, 208Y, 210W, 214F, 215F, 215Y, 219E, 219N, and

-46-

219Q. The 60 RT mutation model consisted of the 26 RT mutation model, plus the 34 next most frequent mutations in d4T resistant samples. These 34 mutations were 20R, 35I, 39A, 43E, 60I, 65R, 122K, 123E, 135T, 162C, 177E, 196E, 200A, 207E, 211K, 228H, 272A, 277K, 286A, 293V, 297K, 329L, 356K, 357T, 358K, 359S, 360T, 371V, 375V, 376A, 386I, 390R, 399D, and 400A.

In this Example, a three layer neural network was employed. The input nodes were used to represent the genotypic mutations. Thus, the 9RT model had 9 input nodes, the 26 RT model had 26 input nodes and the 60 RT model had 60 input nodes. The output nodes were used to represent the degree of resistance to d4T. The hidden nodes were used to determine a suitable model order and achieve good performance. The best architecture for each model was determined by the number of hidden nodes with which the best performance on the independent test data set was achieved. A back-propagation momentum algorithm was also employed. The learning rate is set to 0.01, 0.01, and 0.03 for the 9-model, 26-model, and 60-model respectively. Epochs is set to 10,000, error tolerance is set to 0.0001, and momentum term is set to 0.1 for all three models.

To train and test these neural network models, a total of 2286 samples were used, 188 of which were randomly selected as a test data set. Of the 2286 samples, 1040 of these had d4T IC50 < 3-fold (mean = 1.2), 1246 of these had d4T IC50 > 3-fold (mean = 9.0). Of the 188 test samples (randomly selected from 2286), 92 of these had d4T IC50 < 3-fold (mean = 1.1), and 96 of these had d4T IC50 > 3-fold (mean = 7.7). An optimal solution for each of the models was obtained using the same training and testing data sets. But the complexity of each model was different due to the different number of mutations used, which was affected by the architecture of the neural network model. After each of the neural networks were trained, the test data set was run through each of the networks. The results demonstrated that the 9-mutation model gave a low resistance prediction rate (46%) using the independent test data set and a low concordance rate in the training set (42%). However, the 26- and 60-mutation models could be well trained and also provided a higher prediction rate (80% and 72%, respectively) for resistance (defined as >3-fold increase relative to a sensitive control) using the test data set.

In order to discover which mutations had contributed to this improved prediction, improved sample IS9-26 and IS9-60 were identified by comparing the phenotypic outputs of the 9-model and 26-model, and the 9-model and 60-model on the test set. The corresponding genotypes of the improved samples were collected and analyzed, all extra mutations contained in the improved samples were screened out, and the frequency of each mutation found in IS9-26 and IS9-60 was calculated and compared with that of the mutation being found in the whole samples. All mutations

-47-

with higher difference of two frequencies were identified and considered to play a role in conferring resistance to d4T.

In this example, the threshold frequency was set to 9%. The following mutations were identified from the 9- and 26-models: 41L (44%-79%), 44D (13%-26%), 67N (36%-
5   56%), 70R (21%-30%), 181I (21%-36%), 210W (34%-65%), and 215Y (44%-73%) (Figure 11). The following mutations were identified from the 9- and 60-models: 41L (44%-73%), 67N (36%-56%), 118I (21%-32%), 210W (34%-62%), 211K (49%-59%), and 215Y (44%-74%) (Figure 12). n conclusion, these results show that at least 17 RT mutations (the 8 identified here plus the 9 identified above from the Stanford Database)
10   may confer d4T resistance, including AZT resistance mutations. The results also identified 10 other mutations that may also confer resistance: 184V(36%-42%), 214F(88%-94%), 75A(0.7%-0.6%), 75M(4%-8%), 115F(1%-0.2%), 208Y(13%-21%), 215F(9%-11%), 219E(5%-4%), 219N(4%-11%), and 219Q(12%-16%).
15

Example 7 : A 28-Mutation Neural Network Model That Accurately Predicts Phenotypic Resistance to Lopinavir (LPV)

It has been reported that mutations at 11 codons in HIV-1 protease (10, 20, 24, 46, 53, 54, 63, 71, 82, 84, 90) may be involved in LPV resistance and clinical failure to
20   therapies containing Lopinavir. An optimal set of mutations for the quantitative prediction of LPV resistance has been established using the methods of the present invention and compared with the predictions made by the 11-codon set.

Neural network (NN) models were constructed using 1322 genotyped and phenotyped samples. 80% of these were LPV sensitive (<2.5-fold) and 11% had 'high-level'
25   resistance (>10-fold). 117/1322 samples were randomly selected as a validation set and the remaining 1205 samples used for training. Two NN models were constructed; one based on the 11 previously reported codons (a total of 54 polymorphisms) and a second based on 28 mutations selected by a combination of mutation prevalence analysis and NN pruning techniques.

30   The 28-mutation model gave a high correlation between predicted fold-resistance and actual susceptibility values (r^2=0.88 in the validation and training sets). The 11-codon model gave a similar correlation coefficient for the validation set (r^2=0.84). However, when the predicted values were compared directly with actual fold-resistance, the 28-mutation model was significantly better at predicting LPV
35   resistance compared to the 11-codon model (p<0.001). The proportion of sensitive, low-level and high-level resistance relative to the mutation number per sample was

simulated using the 28-mutation model (n=11x1000). This model demonstrated that samples with as few as 3-4 PI mutations could have high-level LPV resistance.

These results show that LPV resistance can be described by a set of 28 mutations in HIV-1 protease (10I, 18V, 24I, 32I, 33F/M, 43T, 45T, 46I/L, 48V, 53L,
5   54A/S/V, 55R, 58E, 71V, 72Y, 73S/T, 74S, 82A, 84V, 85V, 90M, 95F/L) and that neural network models can be used to accurately quantify LPV resistance based on the genotype.

EXAMPLE 8 : Validation and training of neural network
10   **Development and training of neural networks.** A generic framework for modeling the relationship between genotype and phenotype of HIV-1 drug resistance was developed. This consisted of several phases: establishing a neural network architecture; collecting data; selecting mutations known to correlate with PI resistance, partitioning data; training and testing the system; and statistical analysis. The neural network
15   architecture comprised three 'layers': an input layer (genotypic resistance data); a hidden layer (data processing); and an output layer (predicted phenotypic resistance). The network had I inputs, J hidden units, K output units, and two bias units both of which had an input signal of 1 (i.e., x_0 and z_0) and one bias unit in the input layer which had an input signal of 1 (i.e., x_0). The model was based on 39 mutations associated with
20   PI resistance. These mutations were chosen as they were the most frequently observed genetic polymorphisms in PI resistant samples from our database relative to PI susceptible samples (data not shown). A total of 1015 samples (randomly selected from the database) were used to train the neural network model and 53 randomly selected, independent samples were used as the test data set; an optimal solution for the model
25   was obtained by evaluating the performance of the neural network model on the training and testing data sets. The number of inputs, I, for the model was equal to 39. Each hidden layer, J, contained a plurality of hidden nodes that were adjusted to achieve high predictive performance of the network. The optimal number of hidden nodes was 27 for this model. This was determined empirically. The network had two layers of adaptive
30   weights (w_{ji} and w_{jk}) which are the weight of the jth hidden unit associated with input signal x_i, and the weight of the kth output unit associated with the hidden signal z_j, respectively. The values of these weights were optimised during the training step. The output unit, K, for the model was the predicted phenotypic resistance to the protease inhibitors: indinavir, ritonavir, nelfinavir, saquinavir and amprenavir (defined as >4-fold
35   increase in IC_{50} relative to a sensitive control).
A total of 108 individual, different amino acid changes were used in the search procedure (at a total of 56 unique positions). This was broken down into 39 changes in

-49-

the protease and 69 in the RT (32 for the non-nucleoside RT inhibitors and 37 for the nucleoside analogues). The following mutations, grouped by drug class, were included in the search engine. Protease inhibitors: 10F/I/R/V, 20I/M/R/T, 24I, 30N, 32I, 33F/I/M/V, 36I, 46I/L, 47L, 48V, 50V, 54L/M/V, 71T/V, 73A/C/S, 77I, 82A/F/S/T, 84A/V, 88D/S, 90M. Nucleoside analogues: 41L, 44A/D, 62A, 65R, 67N, 69D/N, 69 insertion, 70R, 74V/I, 75A/I/M/T, 77L, 100I, 115F, 116Y, 118I, 151M, 181C, 184I/T/V, 208Y, 210W, 211K/Q, 215F/Y, 219E/N/Q, 333D/E. NNRTIs: 98G/S, 100I, 101E/I/P/Q, 103N/Q/R/S/T, 106A/I/L, 108I, 179D/E, 181C/I/V, 188C/H/L, 189I, 190A/E/S, 225H, 233V, 236L, 238T.

Validation of resistance mutations using neural networks

The mutation search criteria used for the pattern recognition are extremely comprehensive and constantly updated to include new mutations and polymorphisms that influence phenotypic drug resistance. Although the influence of some mutations on resistance phenotype is relatively straightforward, in many cases (for example protease inhibitor (PI) resistance) there is either a very complex relationship between genetic mutations and the resultant phenotype, or a lack of published information about this relationship. To address this we have trained neural networks to facilitate the identification of new mutations and combinations of mutations that affect drug susceptibility. A neural network was trained using a back propagation learning algorithm using 39 mutations that frequently appear in PI resistant strains as the input layer. 1015 clinical samples with a wide range of PI phenotypic resistance were used as the training set and 53 independent samples were used as a test set. The neural network had a correct rate for predicting resistance to five clinically approved PIs of 87%. Furthermore, there was an excellent correlation between the predicted magnitude of resistance and the actual resistance as determined by phenotypic assay. The overall correlation coefficient (r value) for this analysis was 0.91. Consequently, we used this group of 39 mutations in the search engine to identify clinical samples with PI resistance. In addition, we have also trained neural networks to enable the identification of stavudine (d4T) resistance mutation, as it has frequently been difficult to associate specific RT mutations with d4T resistance.

All references, patents, and patent application cited herein are incorporated by reference in their entirety.

Other embodiments of the invention will be apparent to those skilled in the art from consideration of the specification and practice of the invention disclosed herein. It is intended that the specification and examples be considered as exemplary only, with a true scope and spirit of the invention being indicated by the following claims.

-50-

WHAT IS CLAIMED IS:

1. A method for predicting resistance of a disease to a therapeutic agent comprising:

    (a) providing a trained neural network;

    (b) providing at least one determined genetic sequence from the disease; and

    (c) predicting resistance of the disease to the therapeutic agent using the at least one determined genetic sequence and the trained neural network.

2. The method of claim 1, wherein the disease is chosen pathogens, malignant cells, proliferative cells, and inflammatory cells.

3. The method of claim 2, wherein the pathogen is chosen from disease-producing bacteriums, disease-producing viruses, disease-producing algae, disease-producing fungi, and disease-producing protozoa.

4. The method of claim 3, wherein the pathogen is a disease-producing virus.

5. The method of claim 4, wherein the disease-producing virus is chosen from human immunodeficiency virus type 1, human immunodeficiency virus type 2, herpes simplex virus type 1, herpes simplex virus type 2, human papillomavirus virus, hepatitis B virus, hepatitis C virus, and Epstein-Barr virus.

6. The method of claim 1, wherein the trained neural network is a three-layer feed-forward neural network.

7. The method of claim 6, wherein the three-layer feed forward network comprises:

    (a) a set of input nodes, wherein each member of the set of input nodes corresponds to a mutation in the genome of the pathogen;

    (b) a plurality of hidden nodes; and

    (c) a set of output nodes, wherein each member of the set of output nodes corresponds to a therapeutic agent used to treat the pathogen.

8. The method of claim 1, wherein the predicted resistance is expressed as a fold change in IC50.

9. The method of claim 1 wherein expression levels of the genetic sequence is used.

10. A method for predicting resistance of a disease to a therapeutic agent using a trained neural network comprising:

    (a) providing at least one determined genetic sequence from the disease; and

-51-

    (b) predicting resistance of the disease to the therapeutic agent using the at least one determined genetic sequence and the trained neural network.

11. A method for predicting resistance of a pathogen to a therapeutic agent comprising:

    (a) providing a trained neural network;

    (b) providing a determined genetic sequence from the pathogen; and

    (c) predicting resistance of the pathogen to the therapeutic agent using the determined genetic sequence and the trained neural network.

12. The method of claim 11, wherein the pathogen is chosen from disease-producing bacteriums, disease-producing viruses, disease-producing algae, disease-producing fungi and disease-producing protozoa.

13. The method of claim 12, wherein the pathogen is a disease-producing virus.

14. The method of claim 13, wherein the disease-producing virus is chosen from human immunodeficiency virus type 1, human immunodeficiency virus type 2, herpes simplex virus type 1, herpes simplex virus type 2, human papillomavirus virus, hepatitis B virus, hepatitis C virus, and Epstein-Barr virus.

15. A method for predicting resistance of a pathogen to a therapeutic agent comprising:

    (a) providing a neural network;

    (b) training the neural network on a training data set, wherein each member of the training data set corresponds to a genetic mutation that correlates to a change in therapeutic agent resistance;

    (c) providing a determined genetic sequence from the pathogen; and

    (d) predicting resistance of the pathogen to the therapeutic agent using the determined genetic sequence and the trained neural network.

16. The method of claim 15, wherein the pathogen is chosen from disease-producing bacteriums, disease-producing viruses, disease-producing algae, disease-producing fungi and disease-producing protozoa.

17. The method of claim 16, wherein the pathogen is a disease-producing virus.

18. The method of claim 17, wherein the disease-producing virus is chosen from human immunodeficiency virus type 1, human immunodeficiency virus type 2,

-52-

herpes simplex virus type 1, herpes simplex virus type 2, human papillomavirus virus, hepatitis B virus, hepatitis C virus, and Epstein-Barr virus.

19. The method of claim 15, wherein the neural network is a three-layer feed-forward neural network.

20. The method of claim 19, wherein the three-layer feed forward network comprises:
    (a) a set of input nodes, wherein each member of the set of input nodes corresponds to a mutation in the genome of the pathogen;
    (b) a plurality of hidden nodes; and
    (c) a set of output nodes, wherein each member of the set of output nodes corresponds to a therapeutic agent used to treat the pathogen.

21. A trained neural network capable of predicting resistance of a disease to a therapeutic agent, wherein the trained neural network comprises:
    (a) a set of input nodes, wherein each member of the set of input nodes corresponds to a mutation in the genome of the disease; and
    (b) a set of output nodes, wherein each member of the set of output nodes corresponds to the therapeutic agent used to treat the disease.

22. The trained neural network according to claim 21, wherein the disease is a pathogen.

23. The trained neural network according to claim 22, wherein the pathogen is chosen from a disease-producing bacterium, a disease-producing virus, a disease-producing algae, a disease-producing fungus, and a disease-producing protozoa.

24. A method of designing a therapeutic agent treatment regimen for a patient afflicted with a disease comprising:
    (a) providing a determined genetic sequence from the disease;
    (b) inputting the determined genetic sequence into a trained neural network;
    (c) predicting resistance of the disease to a therapeutic agent using the determined genetic sequence and the trained neural network; and
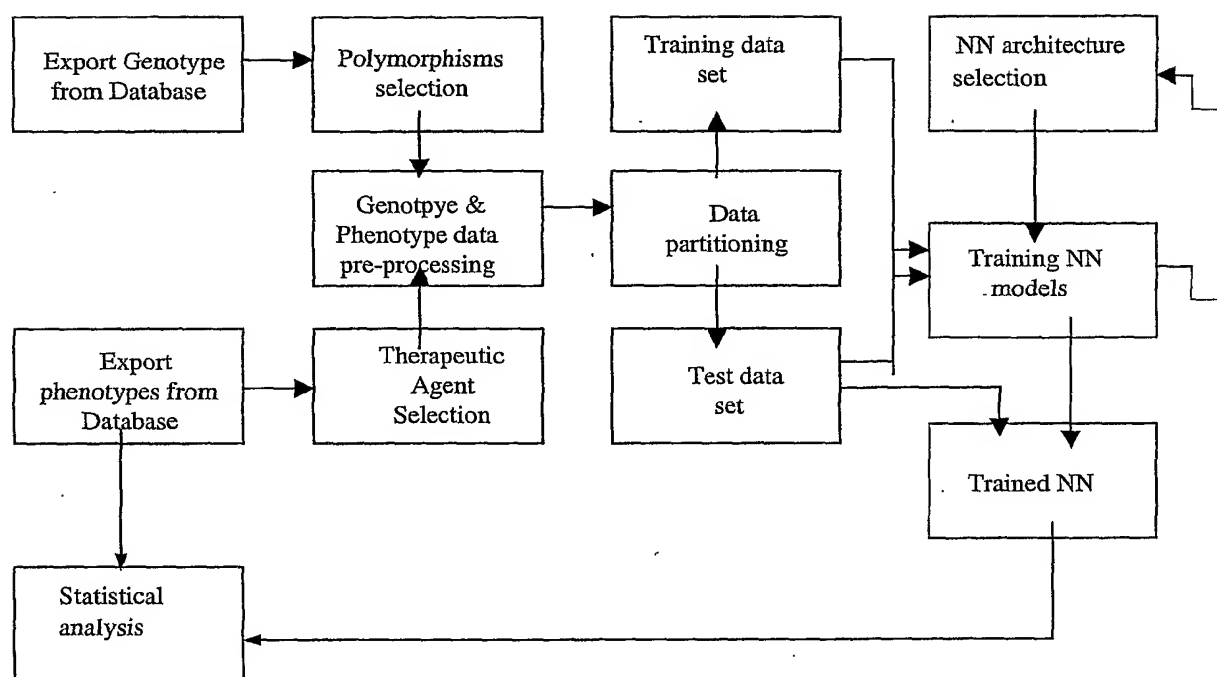   (d) using the predicted drug resistance to design the therapeutic drug treatment regimen to treat the patient afflicted with the disease.

25. The method of claim 24, wherein the disease is chosen from a pathogen and a malignant cell.

-53-

26. The method of claim 25, wherein the pathogen is chosen from a disease-producing bacterium, a disease-producing virus, a disease-producing algae, a disease-producing fungus, and a disease-producing protozoa.

27. The method of claim 26, wherein the pathogen is a disease-producing virus.

28. The method of claim 27, wherein the disease-producing virus is chosen from human immunodeficiency virus type 1, human immunodeficiency virus type 2, herpes simplex virus type 1, herpes simplex virus type 2, human papillomavirus virus, hepatitis B virus, hepatitis C virus, and Epstein-Barr virus.

29. The method of claim 28, wherein the disease-producing virus is the human immunodeficiency virus type 1.

30. A method of predicting the probability of a patient developing a genetic disease comprising:
    (a) providing a trained neural network;
    (b) providing a determined genetic sequence from a patient sample; and
    (c) determining the probability of the patient of developing the genetic disease using the determined genetic sequence and the trained neural network.

31. A method for identifying a new mutation that confers resistance to a therapeutic agent comprising:
    (a) providing a first trained neural network, wherein the number of input nodes for said first trained neural network is equal to the number of mutations known to confer therapeutic resistance to a therapeutic agent;
    (b) providing a second trained neural network, wherein the number of input nodes of said second trained neural network comprises the number of mutations known to confer therapeutic resistance to a therapeutic agent plus at least one additional mutation;
    (c) providing a test data set;
    (d) inputting the test data set into the first and second trained neural networks;
    (e) comparing the output of the first and second trained neural networks to determine whether the additional mutation confers therapeutic drug resistance to a disease.

32. A method for studying therapeutic agent resistance comprising:
    (a) mutating a wild type gene to create a mutant containing a mutation identified using the method of claim 31;

-54-

    (b) culturing the mutant in the presence of a therapeutic agent;

    (c) culturing the wild gene in the presence of the therapeutic agent; and

    (d) comparing the growth of the mutant against the growth of the wild-type.

33. The method of claim 24, wherein a report is created that provides the predicted resistance of the disease to a therapeutic agent, and the report is used by a clinician to design the therapeutic drug treatment regimen to treat the patient afflicted with the disease.

34. A computer-readable medium containing instructions for causing a computer to perform a method for predicting resistance of a disease to a therapeutic agent using a trained neural network, the method comprising:

    receiving at least one determined genetic sequence from the disease; and

    predicting resistance of the disease to the therapeutic agent using the at least one determined genetic sequence and the trained neural network.

35. A computer-readable medium containing a set of program instructions for causing a computer to provide a neural network to perform a method for predicting resistance of a disease to a therapeutic agent, the set of program instructions comprising:

    means for receiving at least one determined genetic sequence from the disease; and

    means for predicting resistance of the disease to the therapeutic agent using the at least one determined genetic sequence and the trained neural network.
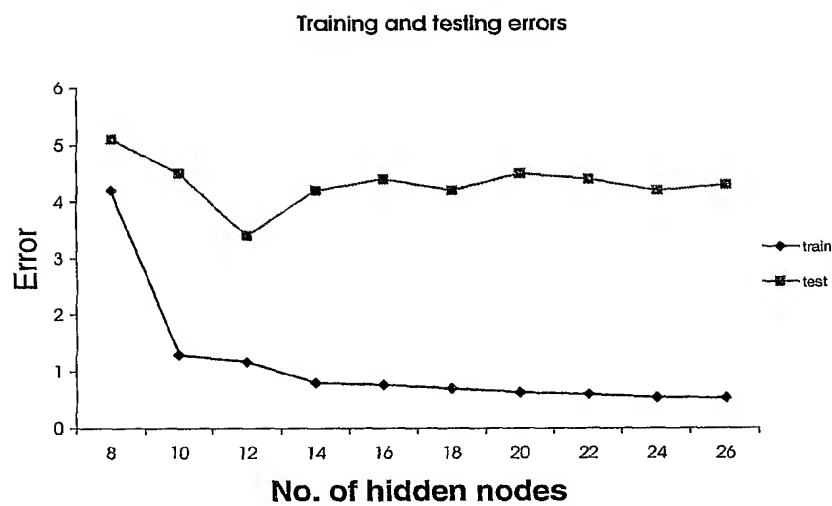
Figure 1

Figure 2

Figure 3

Training and testing errors

Figure 4 (a)

Training and test errors
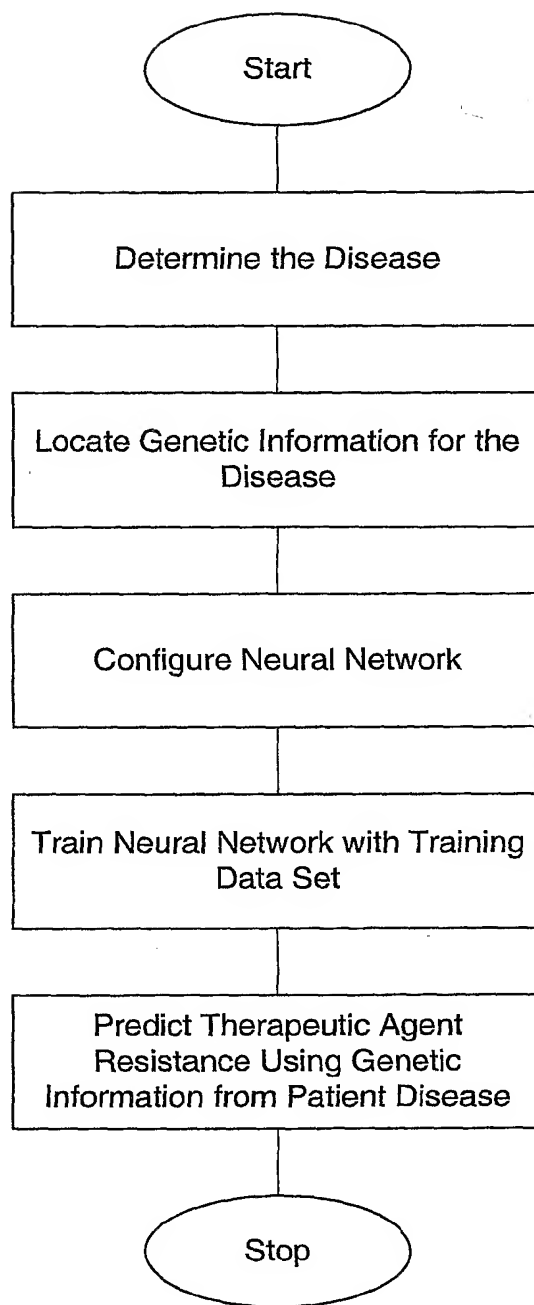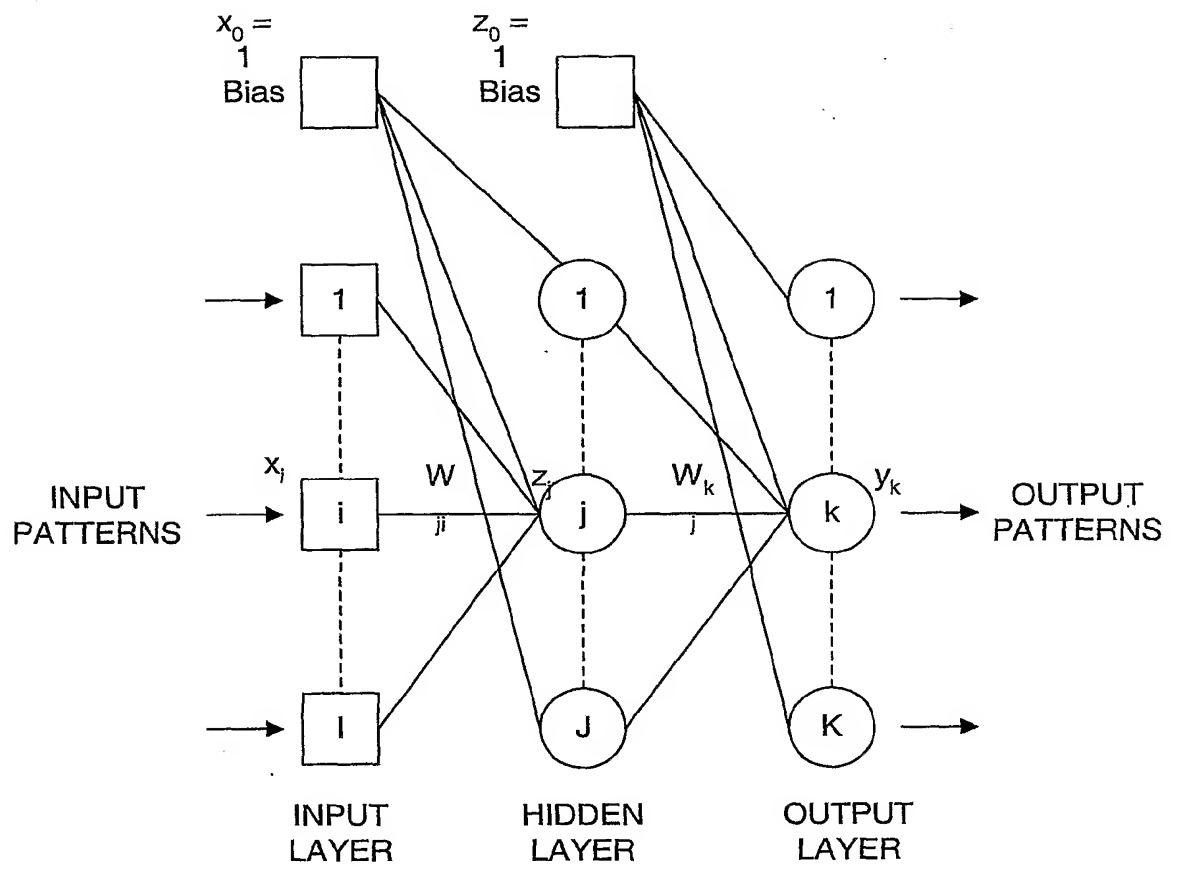
Figure 4 (b)
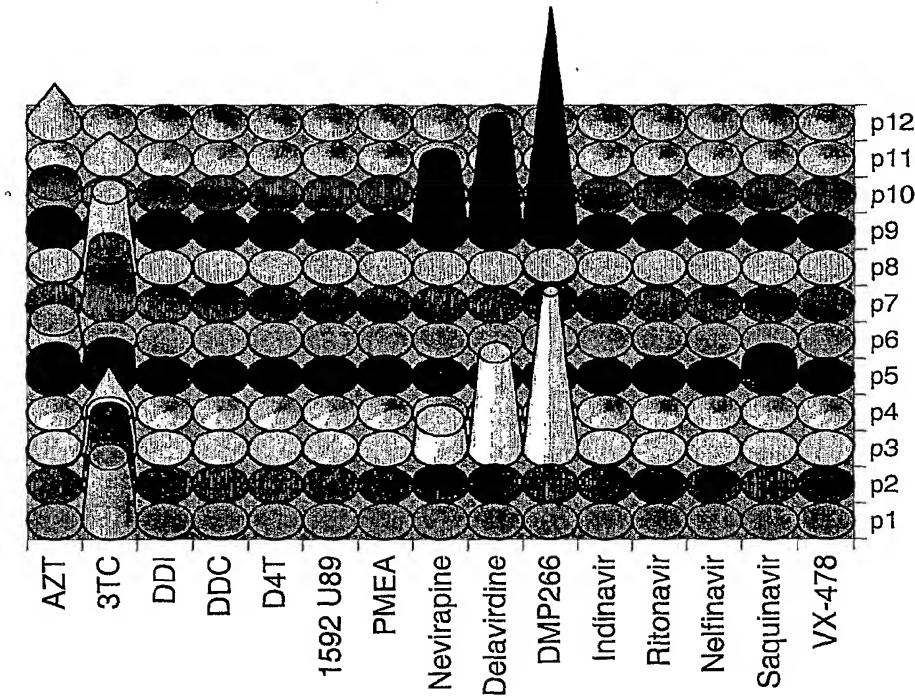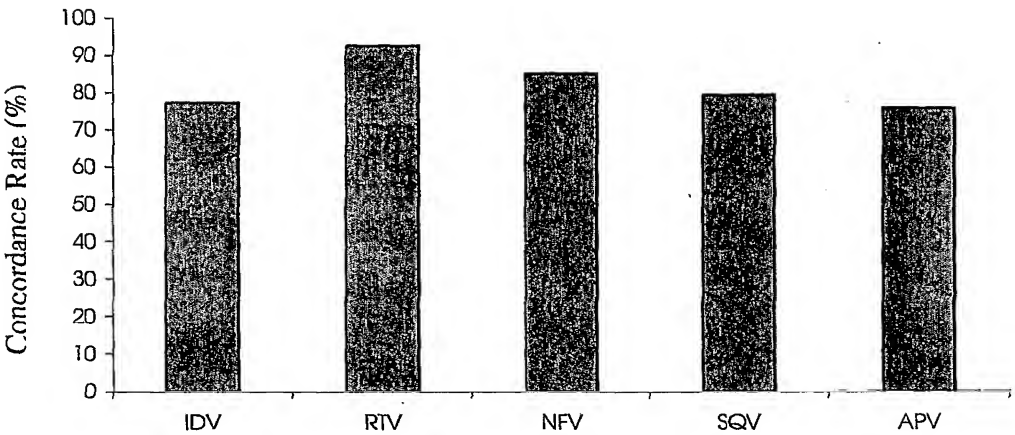
Figure 5

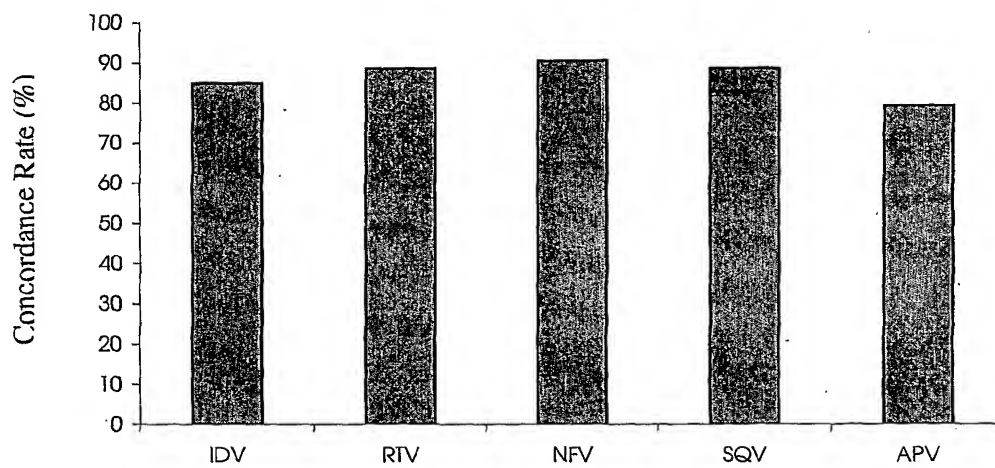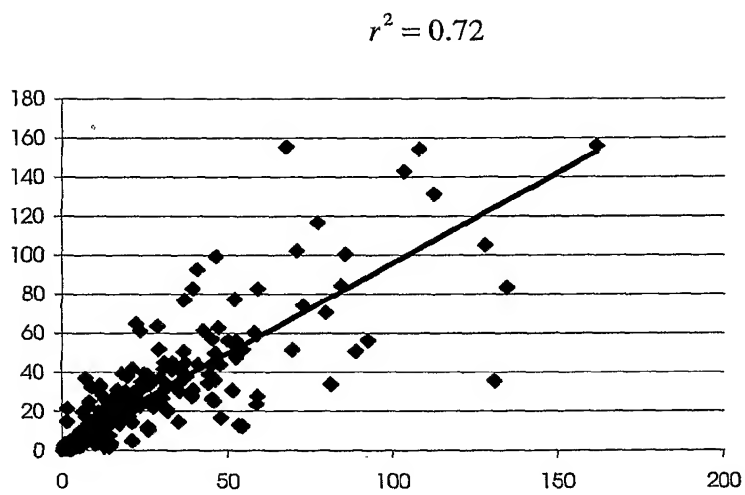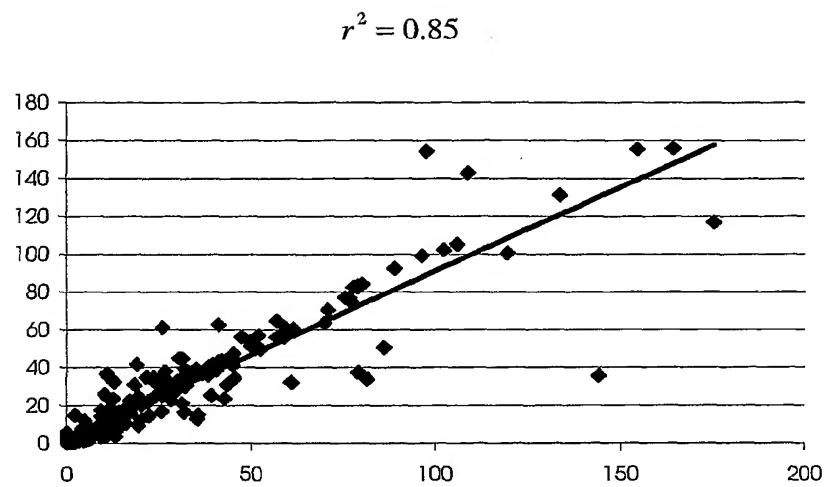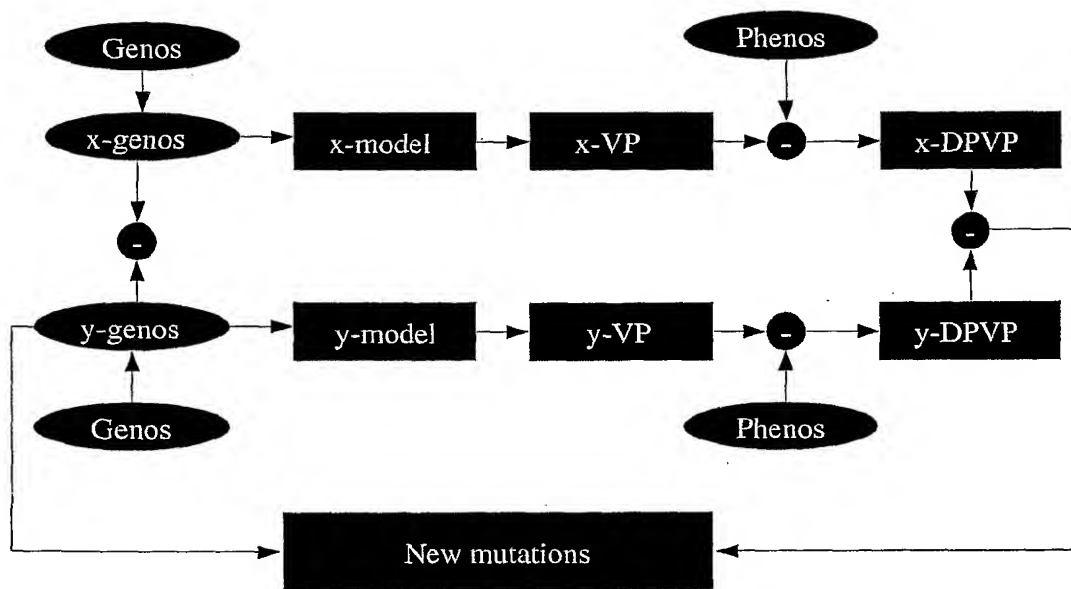
Figure 6

Figure 7

$$r^2 = 0.72$$

Figure 8

$$r^2 = 0.85$$

Figure 9

Figure 10

Figure 11

Figure 12